Factors affecting reliablity of state-space age-structured assessment models

Timothy J. Miller[1,2] Greg Britten[3] Elizabeth N. Brooks[2]

Gavin Fay[4] Alex Hansell[2] Christopher M. Legault[2]

Chengxue Li[2] Brandon Muffley[5] Brian C. Stock[6]

John Wiedenmann[7]

17 April, 2025

[1]corresponding author: timothy.j.miller@noaa.gov

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

[3]Biology Department, Woods Hole Oceanographic Institution, 266 Woods Hole Rd. Woods Hole, MA, USA

[4]Department of Fisheries Oceanography, School for Marine Science and Technology, University of Massachusetts Dartmouth, 836 S Rodney French Boulevard, New Bedford, MA 02740, USA

[5]Mid-Atlantic Fishery Management Council, 800 North State Street, Suite 201, Dover, DE 19901 USA

[6]Institute of Marine Research, Nye Flødevigveien 20, 4817 His, Norway

[7]Department of Ecology, Evolution, and Natural Resources. Rutgers University

keywords: stock assessment, state-space, model selection, bias, convergence, retrospective patterns

Abstract

State-space models are increasingly used for stock assessment, and evaluations of their statistical reliability and best practices for selecting among process error configurations are needed. We simulated 72 operating models that varied fishing pressure and observation error across process errors in recruitment, survival, selectivity, catchability, and/or natural mortality. We fit estimating models with different assumptions on the process error source and whether median natural mortality or a stock-recruit relationship were estimated. Estimating models without a stock-recruit relationship that assumed the correct process error source and median natural mortality had high convergence rates and low bias. Bias was also low under many incorrect process error assumptions when there was contrast in fishing pressure and low observation error. 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 small but were sizable for recruitment when observation error was high. These results help establish the statistical reliability of state space assessment models and pave the way for the next-generation of fisheries stock assessment

Introduction

Application of state-space models in fisheries stock assessment and management has expanded dramatically within International Council for the Exploration of the Sea (ICES), Canada, and the Northeast US (Nielsen and Berg 2014; Cadigan 2016; Pedersen and Berg 2017; Stock and Miller 2021). State-space models latent population characteristics as statistical time series with periodic observations that also may have error due to sampling or other sources of measurement error. Traditional assessment models may use state-space approaches to account for temporal variability in population characteristics (Legault and Restrepo 1999; Methot and Wetzel 2013), but these models treat the annual parameters as penalized fixed effects parameters where the variance parameters controlling the penalties are assumed known (Thorson and Minto 2015). Modern state-space models can estimate the annually varying parameters as random effects with variance parameters estimated using maximum marginal likelihood or corresponding Bayesian approaches. These latter approaches are considered best practice and a recommended for the next generation of stock assessment models (Hoyle et al. 2022; Punt 2023).

State-space stock assessment models, with nonlinear functions of latent parameters and multiple types of observations with varying distributional assumptions, are one of the most complex examples of this analytical approach. Statistical aspects of state-space models and their application within fisheries have been studied extensively, but previous work has focused primarily on linear and Gaussian state-space models (Aeberhard et al. 2018; Auger-Méthé et al. 2021). Therefore, current understanding of the reliability of state-space models does not extend to usage for stock assessment.

As state-space models provide greater flexibility by allowing multiple processes to vary as random effects (Nielsen and Berg 2014; Aeberhard et al. 2018; Stock et al. 2021), one of the most immediate questions regards the implications of mis-specification among alternative sources of process error. Incorrect treatment of population attributes as temporally varying

(Trijoulet et al. 2020; Liljestrand et al. 2024) could lead to misidentification of stock status and biased population estimates, ultimately impacting fisheries management decisions (Legault and Palmer 2016; Szuwalski et al. 2018; Cronin-Fine and Punt 2021). Furthermore, biological, fishery, and observational processes are often confounded in catch-at-age data, which may adversely affect ability to distinguish between true process variability and observational error (Li et al. In review; Punt et al. 2014; Stewart and Monnahan 2017; Cronin-Fine and Punt 2021; Fisch et al. 2023).

Li et al. (2024) conducted a full-factorial simulation-estimation study to assess model reliability when confounding random-effects processes (numbers-at-age, fishery selectivity, and natural mortality) were included. Their results suggest that while state-space models can generally identify sources of process error, overly complex models, even when misspecified (i.e., incorporating process error that did not exist in reality), often performed similarly to correctly specified models, with little to no bias in key management quantities. Similarly, Liljestrand et al. (2024) found little downside in assuming process error in recruitment or selectivity, even when it was absent.

Despite mounting efforts, several limitations remain. First, confounding processes that can be treated as random effects in the model were not thoroughly examined or tested within a simulation-estimation framework. Second, previous studies relied on operating models conditioned on specific fisheries, limiting their generalizability (Li et al. In review; Liljestrand et al. 2024). In particular, the effects of observation error and underlying fishing history have not been fully isolated in simulation study designs, making it challenging to disentangle the interplay between process and observation error magnitudes, as demonstrated in Fisch et al. (2023). Third, explicitly modeling stock-recruit relationships (SRRs) as mechanistic drivers of population dynamics is promising (Fleischman et al. 2013; Du Pontavice et al. 2022), but can reliability of inferences within integrated state-space age-structured models has not been evaluated. Evidence from other studies suggests that when both process and observation errors are unknown, estimating density dependence parameters becomes highly uncertain (Knape 2008; Polansky et al. 2009). In particular, Knape (2008) demonstrated that stronger density dependence becomes increasingly difficult to estimate in the presence of observation error. Therefore, it is crucial to assess whether density dependence mechanisms can be estimated with sufficient precision for use in fisheries management (Auger-Méthé et al. 2016).

In the present study, we conduct a simulation study with operating models (OMs) varying by degree of observation error, 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 SRR was estimated, and whether a constant, or, in some EMs, median, natural mortality is estimated. Given the confounding nature of process errors, developing diagnostic tools to detect model misspecification is of great scientific interest and could aid the next generation of stock assessments (Auger-Méthé et al. 2021). We evaluate whether convergence and Akaike Information Criterion (AIC) can correctly determine the source of process error and the existence of a SRR. We also evaluate when retrospective patterns occur and 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 OMs and EMs in our simulation study (Miller and Stock 2020; Stock and Miller 2021). WHAM is an R package freely available via a github repository and is built on the Template Model Builder package (Kristensen et al. 2016). For this study we used version 1.0.6.9000, commit 77bbd94. WHAM has also been used to configure OMs and EMs for closed loop simulations evaluating index-based assessment methods (Legault et al. 2023) and is currently used or accepted for use in management of numerous fish stocks in the Northeast United States (NEUS) (e.g., NEFSC 2022a, 2022b; NEFSC 2024).

We completed a simulation study with a number of OMs that can be categorized based on where process error random effects were assumed: recruitment (R, assumed present in all models), apparent survival (denoted R+S), natural mortality (R+M), fleet selectivity (R+Sel), or index catchability (R+q). We refer to the (R+S) OMs as modeling apparent survival because on logscale the random effects ($\epsilon_{a,y}$) are additive to the total mortality (F+M) between numbers at age, thus they modify the survival term. However, as Stock and Miller (2021) note, these random effects can be due to events other than mortality, such as immigration, emigration, missreported catch, and other sources of misspecification. For each OM, assumptions about the magnitude of the variance of process errors and observations are required and the values we used were based on a review of the range of estimates from NEUS assessments using WHAM in stocks NEUS.

In total, we configured 72 OMs with alternative assumptions about the source and magnitude of process errors, magnitude of observation error in indices and age composition data, and contrast in fishing pressure over time. We fitted 20 EMs to observations generated from each of 100 simulations where process errors were also simulated. Each EM differed in assumptions about the source of process errors, whether natural mortality (or the median for models with process error in natural mortality) was estimated, and whether a Beverton-Holt SRR was estimated within the EM. Details of each of the OMs and EMs 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 OMs and EMs to simplify interpretation of the study results (Li et al. In review). Simulations and model fitting were all carried out on the University of Massachusetts Green High-Performance Computing Cluster. All code we used to perform the simulation study and summarize results can be found at https://github.com/timjmiller/SSRTWG/tree/main/Project_0/code.

Operating models

Population

The population consists of 10 age classes, ages 1 to 10+, with the last being a plus group that accumulates ages 10 and older. We assume spawning occurs annually 1/4 of the way through the year. The maturity at age was a logistic curve with $a_{50} = 2.89$ and slope $= 0.88$ (Figure S1, top left).

Weight at age was generated with a von Bertalanffy growth function

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

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

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

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

We assumed a Beverton-Holt SRR with constant pre-recruit mortality parameters for all OMs. All post-recruit productivity components are constant in the apparent survival (R+S) and survey catchability (R+Sel) process error OMs. Therefore, steepness and unfished recruitment are also constant over the time period for those OMs (Miller and Brooks 2021). We specified unfished recruitment equal to e^{10} and $F_{\text{MSY}} = F_{40\%} = 0.348$, which equates to a steepness of 0.69 and $a = 0.60$ and $b = 2.4 \times 10^{-5}$ for the Beverton-Holt parameterization

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

(Figure S1, bottom right). For OMs without process errors on natural mortality we fixed the rate at 0.2. For OMs with process errors on natural mortality, the median natural mortality rate was specified to be 0.2.

We used two fishing scenarios for OMs. In the first scenario, the stock experiences overfishing at $2.5F_{\text{MSY}}$ for the first 20 years followed by fishing at F_{MSY} for the last 20 years (denoted $2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$). In the second scenario, the stock is fished at F_{MSY} for the entire time period (40 years). The magnitude of the overfishing assumptions is based on average estimates of overfishing for NEUS groundfish stocks from Wiedenmann et al. (2019) and similar to the approach in Legault et al. (2023).

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

For OMs with time-varying random effects for M , steepness is not constant. However, we used the same a and b parameters as other OMs, which equates to a steepness and R0 at the median of the time series process for M. For OMs with time-varying random effects for fishery selectivity, F_{MSY} is also not constant, but since we use the same F history as other OMs, this corresponds to F_{MSY} at the mean selectivity parameters.

Fleets

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

Indices

Two time series of fishery-independent surveys in numbers are generated 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 ($a_{50} = 5$ and slope $= 1$) and a logistic-normal distribution with no correlation on the multivariate normal scale for the age-composition observations.

Observation Uncertainty

The standard deviation for log-aggregate catch was 0.1. Two levels of observation error variance (high and low) were specified for indices and all age composition observations (both indices and catch). The low uncertainty specification assumed a standard deviation of 0.1 for both series of log-aggregate index observations, 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, the 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) apparent 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 (2x3 levels) and scenarios for 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 as 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 are marginally distributed as $\epsilon_{M,y} \sim \mathrm{N}\left(0, \sigma_M^2\right)$. Uncorrelated random effects were also included on recruitment with $\sigma_R = 0.5$ (hence, we denote these OMs as 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

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 μ_s and any annual random effects marginally distributed as $\epsilon_{s,y} \sim \mathrm{N}\left(0, \sigma_s^2\right)$, where the lower and upper bounds of the parameter (l_s and u_s) can be specified by the user. All selectivity parameters (a_{50} and slope parameters) were bounded by 0 and 10 for all OMs 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 (2x2 levels) and scenarios described above for fishing history (2 levels) and observation error (2 levels) 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 μ_i and any annual random effects marginally distributed as $\epsilon_{i,y} \sim \mathrm{N}\left(0, \sigma_i^2\right)$ where the lower and upper bounds of the catchability (l_i and u_i) can be specified by the user. We assumed bounds of 0 and 1000 for all OMs and EMs. For all OMs and EMs with process errors on catchability, the temporal variation only applies to the first index, which could be interpreted as capturing some unmeasured seasonal process that affects availability to the survey. 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 of the data sets 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 have different assumptions about the source of process error (R+S, R+M, R+Sel, R+q) and whether or not 1) there is temporal autocorrelation, 2) a Beverton-Holt SRR is estimated, and 3) the natural mortality rate (μ_M, the constant or mean on log scale for R+M EMs) is estimated. For simplicity we refer to the derived estimate e^{μ_M} as the median natural mortality rate regardless of whether natural mortality random effects are estimated in the EM.

Subsets of 20 EMs in Table S5 were fit to simulate data sets from each of the OM process error categories. For R and R+S OMs, fitted EMs had matching process error assumptions as well as R+Sel, R+M, and R+q assumptions without autocorrelation. Similarly, For other OM process error categories, we fit EMs with matching process error assumptions as well as other process error types without autocorrelation. 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. However, the variance parameters for the logistic-normal distributions for age composition observations were estimated in the EMs.

Measures of reliability

0.0.1 Convergence

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; Kapur et al. 2025), but given the importance of estimates of uncertainty when using assessment models in management, we estimated probablity of convergence as measured by occurrence of a positive-definite hessian matrix at the optimized negative log-likelihood that could be inverted. We also provide 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. We used the Clopper-Pearson exact method for constructing 95% confidence intervals of the probabilities of convergence (Clopper and Pearson 1934; Thulin 2014).

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 SRR (Beverton-Holt or none).

We also estimated the probability of correctly selecting between EMs with Beverton-Holt SRR assumed and models without the SRR (null model). We made these comparisons between models that otherwise assumed 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 estimation of SRR 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(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 model selection results 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 invertibity of the hessian because optimization could correctly determine an inappropriate process error assumption by estimating variance parameters at the lower bound of zero. Such an optimization could indicate poor convergence but the likelihood would be equivalent to that without the mis-specified random effects and the AIC would be appropriately higher because more (variance) parameters were estimated. All other measures of reliability described below (bias and Mohn's ρ) use these same criteria for inclusion of EM fits in the summarized results.

Bias

For a given model attribute we calculated the relative error

$$\mathrm{RE}\left(\theta_i\right) = \frac{\widehat{\theta}_i - \theta_i}{\theta_i}$$

from fitting a given estimating model to simulated data set i configured for a given OM where $\widehat{\theta}_i$ and θ_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 distribution approach as in Miller and Hyun (2018) and Stock and Miller (2021). We present results for bias in terminal year estimates of SSB and recruitment, Beverton-Holt stock recruit parameters (a and b), and median natural mortality rate. Results for terminal year fishing mortality are were strongly negatively correlated with those for SSB and are provided in the Supplementary Materials.

Mohn's ρ

We calculated Mohn's ρ for SSB, fully-selected fishing mortality, and recruitment for each EM (Mohn 1999). We fit 7 peels for each EM and calculated median 95% confidence intervals for Mohn's ρ using the same methods as that for relative bias.

Results

Convergence performance

For R and R+S OMs, convergence generally declined for most EMs when the median natural mortality rate was estimated and/or the Beverton-Holt SRR was estimated even when the process error assumptions of the EMs and OMs matched (Figure 1, A). When there was high observation error and constant fishing pressure ($F = F_{\mathrm{MSY}}$ for all 40 years), convergence was poor for of all EM process error configurations other than R EMs when fitted to R OMs ($\sigma_{2+} = 0$) regardless of whether median natural mortality and SRRs 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 when median natural mortality and SRRs were estimated. R+S EMs fit to R OMs exhibited poor convergence regardless of whether natural mortality or a SRR 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 error except those where median natural mortality and/or SRRs were estimated.

Convergence of all EMs fitted to R+M OMs was highest 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 regardless of autocorrelation of the simulated process errors. R+S EMs fitted to data generated from R+M OMs always converged poorly whether or not median natural mortality and the Beverton-Holt SRR were estimated.

The R+S EMs, in particular, had poor convergence when fit to data generated from R+Sel OMs with lower selectivity process error variability or higher observation error (Figure 1, C). R+Sel EMs generally converged better than other EMs for R+Sel OMs with higher process error variability, lower observation error, and contrast in fishing pressure regardless of whether median natural mortality or a SRR was estimated.

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

We found a wide range of maximum absolute values of gradients for models that converged (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 SRR produced the lowest gradient values.

AIC performance for process error structure

Marginal AIC accurately determined the correct process error assumptions in EMs when data were generated from R and R+S OMs, regardless of whether median natural mortality or a SRR was estimated (Figure 2, A). Attempting to estimate median natural mortality or a SRR separately had a negligible effect on the accuracy of determining the correct process error assumption. When both were estimated, there was a noticeable reduction in accuracy when OMs had a constant fishing pressure, low observation error, 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). Of these OMs, estimating the median natural mortality rate only reduced the accuracy of AIC when natural mortality process errors were independent and fishing pressure was constant. For OMs with poor model selection accuracy, AIC most frequently selected EMs with process errors in catchability (R+q) or selectivity (R+Sel). Selection of R+S EMs was generally unlikely.

Marginal AIC most accurately determined the correct source of process error and correlation structure for R+Sel OMs with low observation error (Figures 2, C). When there was low variability in selectivity process errors and high observation error, R+q or R+S EMs were more likely to have the best AIC. Whether median natural morality or SRRs 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 R+q OMs with low variability in catchability process errors and high observation error had the least model selection accuracy. However, for these OMs, the marginal AIC accurately determined the correct source of process error (but not correlation structure) except when OMs assumed a constant fishing pressure and EMs estimated both median natural morality and the SRR.

AIC performance for the stock-recruit relationship

Our comparisons of model performance conditioned on assuming the true process error configuration is known (EM and OM process error types match) and we focus on results where the EMs assume median natural mortality is known because there was little difference in results when the EMs estimated this parameter. Broadly, we found generally poor accuracy of AIC in selecting models assuming a Beverton-Holt SRR over the null model without an SRR for all OMs. However, we also found increased accuracy of AIC in determining the Beverton-Holt SRR when the simulated population exhibited greater variation in spawning biomass for nearly every OM (Figure 3).

With R and R+S process error assumptions, probability of lowest AIC for the B-H SRR as a function of SSB variability were greatest for OMs with contrast in fishing pressure and lower process variability in recruitment (Figure 3, A). The largest variation in SSB occurred in OMs with larger recruitment variability ($\sigma_R = 1.5$; Figure 3, A, right column group), but the same high AIC accuracy was achieved for OMs with lower recruitment variability at lower levels of SSB variation. The level of observation error had little effect on AIC accuracy.

For R+M OMs, probability of lowest AIC for the Beverton-Holt SRR increased steeply with variation in SSB whether it was induced by contrast in fishing or variation in natural mortality process error. (Figure 3, B). There was little difference in AIC accuracy whether the natural mortality process errors were correleted and, similar to R+S OMs, there was also little effect due to level of observation error.

For R+Sel OMs, contrast in fishing pressure over time was the primary source of variation in SSB and these are the OMs where AIC accuracy for the Beverton-Holt SRR was greatest (Figure 3, C). There was little effect of variability or correlation of selectivity process errors or the level of observation error on AIC accuracy.

Like the R+Sel OMs, the greatest accuracy for AIC in selecting the Beverton-Holt SRR 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). There was also little effect of variability or correlation of catchability process errors or the level of observation error on AIC accuracy.

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 error (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 error (Figure 4, A, second row).

For R+S OMs, the EMs with matching process error assumptions generally produced unbiased estimation of terminal SSB except when median natural mortality was estimated and there was high observation error. In R+S OMs with low observation error, EMs with incorrect process error assumptions typically provided biased estimation of terminal year SSB. Estimating the Beverton-Holt SRR 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 error and contrast in fishing motality 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 error and constant fishing pressure over time. It also increased bias in SSB estimation for many R+M OMs. Like R and R+S OMs, estimating a SRR had little discernible effect on SSB bias.

For R+Sel OMs, there was no evidence of bias for any EMs when variability in selectivity process error and observation error was low, and with contrast in fishing mortality (Figure 4, C). The largest bias occurred for any EMs that estimated median natural mortality when the OMs had high observation error, constant fishing pressure, and greater variability in selectivity process errors ($\sigma_{\text{Sel}} = 0.5$) or low selectivity process errors ($\sigma_{\text{Sel}} = 0.1$) and low observation error. However, there was no evidence of SSB bias for correctly specified R+Sel EMs when observation error was low and variation in selectivity process errors was larger, whether median natural mortality was estimated or not (Figure 4, C, third row). We only observed an effect of estimating the Beverton-Holt SRR for R+Sel OMs that had high observation error and contrast in fishing pressure where estimating the SRR produced less biased SSB estimation for many EMs (Figure 4, C, top row).

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

For all OM process error types, relative errors in terminal year recruitment were generally more variable than SSB, but effects of R and R+S OM and EM attributes on bias (i.e, negative or positive or none) were similar (Figure S6, A). Furthermore, for EM configurations where bias in terminal SSB was evident, median relative errors in recruitment often indicated stronger bias in recruitment of the same sign.

Beverton-Holt parameters

Across all OMs, there was generally better accuracy (less bias and/or lower variability) for estimation of the Beverton Holt a parameter than the b parameter. In R and R+S OMs, EMs with the correct assumptions about process errors provided the least biased estimation of Beverton-Holt SRR parameters when there was a change in fishing pressure over time and lower variability of recruitment process errors, but there was little effect of estimating median natural mortality and a small increase in bias for those OMs that had high observation error (Figure S7, A). For other R and R+S OMs, estimating natural mortality often resulted in less biased estimation of SRR parameters. There was generally large variability in relative errors of the SRR 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.

In R+M OMs, the most accurate estimation of SRR parameters for all EM process error assumptions occurred when there was a change in fishing pressure, greater variability in natural mortality process errors, and lower observation error (Figure S7, B). Relative to the R, and R+S OMs, there was even less effect of estimating median natural mortality on estimation bias for the SRR parameters.

Bias for SRR parameters was large and variability in relative errors was greatest for most EMs fit to R+Sel OMs with constant fishing pressure (Figure S7, C). Less bias in parameter estimation occurred for OMs with a change in fishing pressure and the best accuracy occurred for those OMs that had low observation error and more variable and uncorrelated selectivity process errors, and when the EMs hadd with the correct process error assumption. There was little effect of estimating natural mortality on relative errors for SRR parameters.

Like R+Sel OMs, relative errors in SRR parameters for R+q OMs were more accurate for most EM process error types when OMs had contrast in fishing pressure and lower observation error (Figure S7, D). However, the best accuracy occurred for those OMs that had lower variability in catchability process errors. The worst accuracy of SRR parameter estimation regardless of EM type occurred when R+q OMs had low observation error and constant fishing pressure (Figure S7, D, fourth row).

Median natural mortality rate

Across all OMs and EMs there was little effect of estimating SRRs on the bias in estimation of median natural mortality (Figure S8). Median natural mortality rate was estimated accurately by all EM process error types for all R OMs except those with high observation error and constant fishing pressure, in which case relative errors were high (Figure S8, A,

$\sigma_{2+} = 0$). For R+S OMs estimation of median natural mortality rate was most accurate when observation error was low and there was contrast in fishing pressure and the EM process error type was correct.

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 error (Figure S8, 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.

All EM process error types accurately estimated median natural mortality rate for R+Sel OMs that had contrast in fishing pressure, low observation error, and low selectivity process error variability (Figure S8, 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 error was high and fishing pressure was constant.

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 error and low catchability process error variability (Figure S8, D). Most EM process error types produced biased estimation of median natural mortality when R+q OMs had high observaiton error and constant fishing pressure.

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 error and low catchability process error variability (Figure S8, D). Most EM process error types produced biased estimation of median natural mortality when R+q OMs had high observation error and constant fishing pressure.

Mohn's ρ

Mohn's ρ 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 SRRs were estimated (Figure 5, A). The strongest retrospective patterns (highest absolute Mohn's ρ values) occurred in OMs with the largest apparent survival process error variability, high observation error, and contrast in fishing pressure, but only for EMs with the incorrect process error type and where median natural mortality rate was assumed known (median ρ was approximately -0.15). For R+M, R+Sel, and R+q OMs, Mohn's ρ 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 SRRs on Mohn's ρ for any EM process error types.

Mohn's ρ 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 SRRs were estimated (Figure S10, A). However, R and R+S OMs with greater recruitment process variability and higher observation error had median Mohn's ρ for recruitment greater than zero for most EMs even when the EM process error type was correct. In R+S OMs with lower observation error, EMs with the correct process error type exhibited better median Mohn's ρ close to 0 than EMs with the incorrect process error type. For R+M, R+Sel, and R+q OMs, results for Mohn's ρ for recruitment are similar to those for SSB, but the range in median values and variation in Mohn's ρ values for a given OM are generally larger for recruitment (Figure S10, B-D).

Discussion

Convergence

Analyses of model convergence across simulations can be useful for understanding the utility of alternative convergence criteria used in applications to real data for directing the practitioner to more appropriate random effects configurations. It is common during the assessment model fitting process to check that the maximum absolute gradient component is less than some threshold prior to inspecting the Hessian of the optimized likelihood for invertibility (Carvalho et al. 2021). However, there is no accepted standard for the gradient threshold (e.g., Lee et al. 2011; Hurtado-Ferro et al. 2014; Rudd and Thorson 2018) and some thresholds would exclude models that in fact have an invertible Hessian. We found the Hessian at the optimized log-likelihood can often be invertible when the maximum absolute gradient was much larger than what would perceived to be a sensible threshold.

Li et al. (2024) found that convergence rate could be a useful diagnostic especially for separating the correct model from overly complex models. However, the criteria for convergence used in their study may also lead to limited ability to distinguish the correct model from overly simplistic models, a pattern that was also noted by Liljestrand et al. (2024) in which one process error may absorb all sources of process error when the magnitude of other process errors are low.

Often poor convergence result when parameter estimates are at their bounds (Carvalho et al. 2021), and this also applies to variance parameters for random effects with state-space assessment models. Even when Hessian is invertible, parameters that are poorly informed will have extremely large variance estimates. This further inspection can lead to a more appropriate and often more parsimonious model configuration where the problematic parameters are not estimated. For example, process error variance parameters that are estimated close to 0 indicates that the random effects are estimated to have little or no variability and removing these process errors is warranted. Generally, our results suggest we can expect lower probability of convergence of state-space assessment models when estimating natural mortality or SRRs because of the difficulty distinguishing these parameters from others being estimated in assessment model with data that are typically available. Our experiments did not aim to emulate the practitioner decision process in developing model configurations (e.g. removing a source of process error and refitting the model when process error variance parameters were estimated close to 0). Evaluating the efficacy of such a decision process when applying Ems might be important in closed loop simulations (e.g. MSE) aimed at quantifying management performance.

A factor affecting the convergence criteria, particularly for maximum likelihood estimation of models with random effects, is numerical accuracy. All optimizations performed in these simulations are of the Laplace approximation of the marginal likelihood and, therefore, gradients and Hessians are also with respect to this approximation (see TMB::sdreport in the Template Model Builder package). Functionality within the Template Model Builder package exists (i.e., TMB::checkConsistency) to check the validity of the Laplace approximation and the utility of this as a diagnostic for state-space assessment models should be explored further. Furthermore, numerical methods are used to calculate and invert the Hessian for variance estimation for models with random effects. Along with our results, the potential lack of accuracy imposed by these approximations, suggests at least investigating whether the Hessian is positive definite when the calculated absolute gradients are not terribly large.

AIC

Of the OM process error configurations we considered, we found AIC to be accurate for selecting models with process errors on recruitment and apparent 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; a similar result was reported by Liljestrand et al.

(2024). 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 lower observation error.

Across all OM process error configurations, AIC performed poorly in identifying that the presence of the Beverton-Holt SRR in the OM unless there was contrast in fishing pressure possibly in combination with other factors such as lower variability in recruitment process errors (in R and R+S models) or greater variation in natural mortality process errors (for R+M OMs, Fig. 3). As such, properly accounting for process error in natural mortality could be important (Li et al. 2024) when evaluating SRRs in state-space models. Curiously, we did not find a marked effect of the level of observation error on ability to detect the SRR, but it is possible that AIC would perform better if observations have even lower uncertainty than we considered.

Although we did not compare models with alternative SRRs (e.g., Ricker and Beverton-Holt), we do not expect AIC to perform any better distinguishing between relationships. Our finding that AIC tended to choose simpler recruitment models in most cases contrasts with the noted bias in AIC for more complex models (Shibata 1976; Katz 1981; Kass and Raftery 1995), but, whereas those findings apply to more the much more common comparison of models that are fit to raw and independent observations, here we are comparing state-space models which account for observation error and estimate process errors in latent variables.

Our results comport with those of de Valpine and Hastings (2002) who found AIC could not distinguish among state-space SRRs that were fit just to SSB and recruitment observations (i.e., not an assessment model). Similarly, Britten et al. found AIC could not reliably distinguish alternative environmental effects on SRR parameters. However, Miller et al. (2016) did find AIC to prefer a SRR with environmental effects when applied to data for the SNEMA yellowtail flounder stock and AIC also selected an environmental covariate on a SRR for the most recent stock assessment of Georges Bank yellowtail flounder (NEFSC 2025). Both of these yellowtail flounder stocks have large changes in stock size and the values of environmental covariates over time. Additionally, this species is well-observed by the bottom trawl survey that is used for an index in assessment models.

Bias

As expected, bias in all parameters and assessment output was generally improved with lower observation error. Estimation of SRR parameters was reliable in ideal scenarios of low observation error and contrast in fishing for some R+Sel and R+M OMs, but generally estimation was biased and(or) highly variable. We found substantial bias in estimated SRR parameters in R and R+S OMs particularly with high variability in recruitment and apparent survival process errors, suggesting that practitioners should be cautious when models without the SRR suggest this to be the case.

On the other hand, estimation of median natural mortality was reliable in many OM scenarios with contrast in fishing pressure and in some OMs EMs that also estimated the SRR parameters bias for those parameters was improved. Conversely, for some R+Sel and R+q OMs where there was bias in natural mortality due to high observation error, estimating the SRR reduced the bias in median natural mortality rate. However, estimating median natural mortality did cause poor accuracy in SSB estimation in many OMs without contrast in fishing pressure over time and with higher observation error. Thus, estimating median natural mortality should be approached with caution in state-space assessment models, particularly given its significant impact on determination of reference point and stock status (Li et al. 2024).

Retrospective patterns

Incorrect EM process error assumptions did not produce strong retrospective patterns for SSB for any OMs regardless of whether median natural mortality or a SRR was estimated, but some weak retrospective patterns occur when observation error was high and there was contrast in fishing pressure. However, retrospective patterns tended to be more variable for recruitment and were sometimes large even when the EM was 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, management quantities such as biological reference points, and projections may be beneficial (Brooks and Legault 2016).

The general lack of retrospective patterns with mis-specified process errors is perhaps to be expected. Retrospective patterns are often induced in simulation studies by rapid changes in a quantity such as index catchability, natural mortality, or perceived catch during years toward the end of the time series (Legault 2009; Miller and Legault 2017; Huynh et al. 2022; Breivik et al. 2023). In our simulations, the process errors changing over time may have trends in particular simulations, particularly when strong autocorrelation is imposed, but the random effects have no trend on average across simulations. Szuwalski et al. (2018) and lietal24 also found relatively small retrospective patterns when the source of mis-specification was temporal variation in demography attributes. Indeed, it is common for the flexibility provided by temporal random effects to reduce retrospective patterns (Miller et al. 2018; Stock et al. 2021; Stock and Miller 2021), though it does not necessarily indicate a more accurate assessment model (Perretti et al. 2020; Li et al. 2024; Liljestrand et al. 2024). Our results together with the existing literature seem to suggest that when a strong retrospective pattern is observed in an assessment it is more likely to be due to a mis-specification of a rapid shift in some model attribute rather than whether a particular process is assumed to be randomly varying temporally.

Conclusions

Our simulation study examined the importance of several factors for reliable inferences from state-space age-structured assessment models. Across all factors, AIC accurately distinguished models with process errors on recruitment only or on recruitment and apparent survival. Accuracy for other process error types required a strong signal (high process variability) with low noise (low observation uncertainty). Therefore, we expect practitioners will find R+S configurations to provide satisfactory diagnostics across a range of life history and data quality scenarios. Contrast in fishing pressure was consistently an important factor across all measures of reliability we examined. AIC generally performed poorly for selecting the SRR but performance was improved with low recruitment variability and contrast in fishing pressure. However, some bias in estimation in at least one of the SRR parameters existed in nearly all OM-EM combinations. Because bias in terminal SSB and retrospective patterns were indifferent to whether or not the SRR was estimated, and convergence was slightly better without the SRR, a sensible default would be to fit models without an assumed SRR.

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.

Auger-Méthé, M., Field, C., Albertsen, C.M., Derocher, A.E., Lewis, M.A., Jonsen, I.D., and Mills Flemming, J. 2016. State-space models' dirty little secrets: Even simple linear gaussian models can have estimation problems. Scientific reports **6**(1): 26677. doi:10.1038/srep26677.

Auger-Méthé, M., Newman, K., Cole, D., Empacher, F., Gryba, R., King, A.A., Leos-Barajas, V., Mills Flemming, J., Nielsen, A., Petris, G., and others. 2021. A guide to state–space modeling of ecological time series. Ecological Monographs **91**(4): e01470. doi:10.1002/ecm.1470.

Breivik, O.N., Aldrin, M., Fuglebakk, E., and Nielsen, A. 2023. Detecting significant retrospective patterns in state space fish stock assessment. Canadian Journal of Fisheries and Aquatic Sciences **80**(9): 1509–1518. doi:10.1139/cjfas-2022-0250.

Britten, G., Brooks, E.N., and Miller, T.J. Reliability of stock-recruitment function estimation in state space assessment models.

Brooks, E.N., and Legault, C.M. 2016. Retrospective forecasting – evaluating performance of stock projections for new england groundfish stocks. Canadian Journal of Fisheries and Aquatic Sciences **73**(6): 935–950. doi:10.1139/cjfas-2015-0163.

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.

Clopper, C.J., and Pearson, E.S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26**(4): 404–413. doi:10.1093/biomet/26.4.404.

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.

Cronin-Fine, L., and Punt, A.E. 2021. Modeling time-varying selectivity in size-structured assessment models. Fisheries Research **239**: 105927. Elsevier.

de Valpine, P., and Hastings, A. 2002. Fitting population models incorporating process noise and observation error. Ecological Monographs **72**(1): 57–76.

Du Pontavice, H., Miller, T.J., Stock, B.C., Chen, Z., and Saba, V.S. 2022. Ocean model-based covariates improve a marine fish stock assessment when observations are limited. ICES Journal of Marine Science **79**(4): 1259–1273. doi:10.1093/icesjms/fsac050.

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. Oxford University Press.

Fleischman, S.J., Catalano, M.J., Clark, R.A., and Bernard, D.R. 2013. An age-structured state-space stock–recruit model for pacific salmon (oncorhynchus spp.). Canadian Journal of Fisheries and Aquatic Sciences **70**(3): 401–414. NRC Research Press.

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.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., McGilliard, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-Pre, K.A., Whitten, A.R., and Punt, A.E. 2014. Looking in the rear-view mirror: Bias and retrospective patterns in integrated, age-structured stock assessment models.

ICES Journal of Marine Science **72**(1): 99–110. doi:10.1093/icesjms/fsu198.

Huynh, Q.C., Legault, C.M., Hordyk, A.R., and Carruthers, T.R. 2022. A closed-loop simulation framework and indicator approach for evaluating impacts of retrospective patterns in stock assessments. ICES Journal of Marine Science **79**(7): 2003–2016. doi:10.1093/icesjms/fsac066.

Kapur, M.S., Ducharme-Barth, N., Oshima, M., and Carvalho, F. 2025. Good practices, trade-offs, and precautions for model diagnostics in integrated stock assessments. Fisheries Research **281**: 107206. doi:10.1016/j.fishres.2024.107206.

Kass, R.E., and Raftery, A.E. 1995. Bayes factors. Journal of the American Statistical Association **90**(430): 773–795. doi:10.1080/01621459.1995.10476572.

Katz, R.W. 1981. On some criteria for estimating the order of a Markov chain. Technometrics **23**(3): 243–249. doi:10.1080/00401706.1981.10486293.

Knape, J. 2008. Estimability of density dependence in models of time series data. Ecology **89**(11): 2994–3000. doi:10.1890/08-0071.1.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. 2016. TMB: Automatic differentiation and Laplace approximation. Journal of Statistical Software **70**(5): 1–21.

Lee, H.-H., Maunder, M.N., Piner, K.R., and Methot, R.D. 2011. Estimating natural mortality within a fisheries stock assessment model: An evaluation using simulation analysis based on twelve stock assessments. Fisheries Research **109**(1): 89–94. doi:10.1016/j.fishres.2011.01.021.

Legault, C.M. 2009. Report of the retrospective working group, 14–16 january 2008. US Department of Commerce Northeast Fisheries Science Center Reference Document 09-01. US Department of Commerce Northeast Fisheries Science Center. Woods Hole, MA.

Legault, C.M., and Palmer, M.C. 2016. In what direction should the fishing mortality target change when natural mortality increases within an assessment? Canadian Journal of Fisheries and Aquatic Sciences **73**(3): 349–357. NRC Research Press.

Legault, C.M., and Restrepo, V.R. 1999. A flexible forward age-structured assessment program. Col. Vol. Sci. Pap. ICCAT **49**(2): 246–253.

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., Berger, A.M., Goethel, D.R., Langseth, B.J., Schueller, A.M., and Miller, T.J. In review. Random effects on numbers-at-age transitions implictly 2 account for movement dynamics and improve 3 performance within a state-space stock assessment. Canadian Journal of Fisheries and Aquatic Sciences.

Li, C., Deroba, J.J., Miller, T.J., Legault, C.M., and Perretti, C. In reviewIn 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.

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. Elsevier.

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.

Methot, R.D., and Wetzel, C.R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research **142**: 86–99. doi:10.1016/j.fishres.2012.10.012.

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. doi:10.1139/cjfas-2015-0339.

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. Canadian Journal of Fisheries and Aquatic Sciences **75**(5): 691–703. doi:10.1139/cjfas-2017-0035.

Miller, T.J., and Legault, C.M. 2017. Statistical behavior of retrospective patterns and their effects on estimation of stock and harvest status. Fisheries Research **186**: 109–120. doi:10.1016/j.fishres.2016.08.002.

Miller, T.J., O'Brien, L., and Fratantoni, P.S. 2018. Temporal and environmental variation in growth and maturity and effects on management reference points of Georges Bank Atlantic cod. Canadian Journal of Fisheries and Aquatic Sciences **75**(12): 2159–2171. doi:10.1139/cjfas-2017-0124.

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.

NEFSC. 2022a. Final report of the haddock research track assessment working group. Available at https://s3.us-east-1.amazonaws.com/nefmc.org/14b_EGB_Research_Track_Haddock_WG_F

NEFSC. 2022b. Report of the american plaice research track working group. Available at https://s3.us-east-1.amazonaws.com/nefmc.org/2_American-Plaice-WG-Report.pdf.

NEFSC. 2024. Butterfish research track assessment report. US Dept Commer Northeast Fish Sci Cent Ref Doc. 24-03; 191 p.

NEFSC. 2025. Yellowttail flounder research track working group report. Available at https://d23h0vhsm26o6d.cloudfront.net/10c.-Yellowtail-Flounder-RT-WG-Report.pdf.

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.

Perretti, C.T., Deroba, J.J., and Legault, C.M. 2020. Simulation testing methods for estimating misreported catch in a state-space stock assessment model. ICES Journal of Marine Science **77**(3): 911–920. doi:10.1093/icesjms/fsaa034.

Polansky, L., De Valpine, P., Lloyd-Smith, J.O., and Getz, W.M. 2009. Likelihood ridges and multimodality in population growth rate models. Ecology **90**(8): 2313–2320. doi:10.1890/08-1461.1.

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.

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

Rudd, M.B., and Thorson, J.T. 2018. Accounting for variable recruitment and fishing mortality in length-based stock assessments for data-limited fisheries. Canadian Journal of Fisheries and Aquatic Sciences **75**(7): 1019–1035. doi:10.1139/cjfas-2017-0143.

Shibata, R. 1976. Selection of the order of an autoregressive model by akaike's information criterion. Biometrika **63**(1): 117–126. doi:10.1093/biomet/63.1.117.

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

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.

Stock, B.C., Xu, H., Miller, T.J., Thorson, J.T., and Nye, J.A. 2021. Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for Southern New England-Mid Atlantic yellowtail flounder. Fisheries Research **237**: 105873.

Szuwalski, C.S., Ianelli, J.N., and Punt, A.E. 2018. Reducing retrospective patterns in stock assessment and impacts on management performance. ICES Journal of Marine Science **75**(2): 596–609. doi:10.1093/icesjms/fsx159.

Thorson, J.T., and Minto, C. 2015. Mixed effects: A unifying framework for statistical modelling in fisheries biology. ICES Journal of Marine Science **72**(5): 1245–1256. doi:10.1093/icesjms/fsu213.

Thulin, M. 2014. The cost of using exact confidence intervals for a binomial proportion. Electronic Journal of Statistics **8**(1): 817–840. doi:10.1214/14-EJS909.

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. Wiley Online Library.

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.

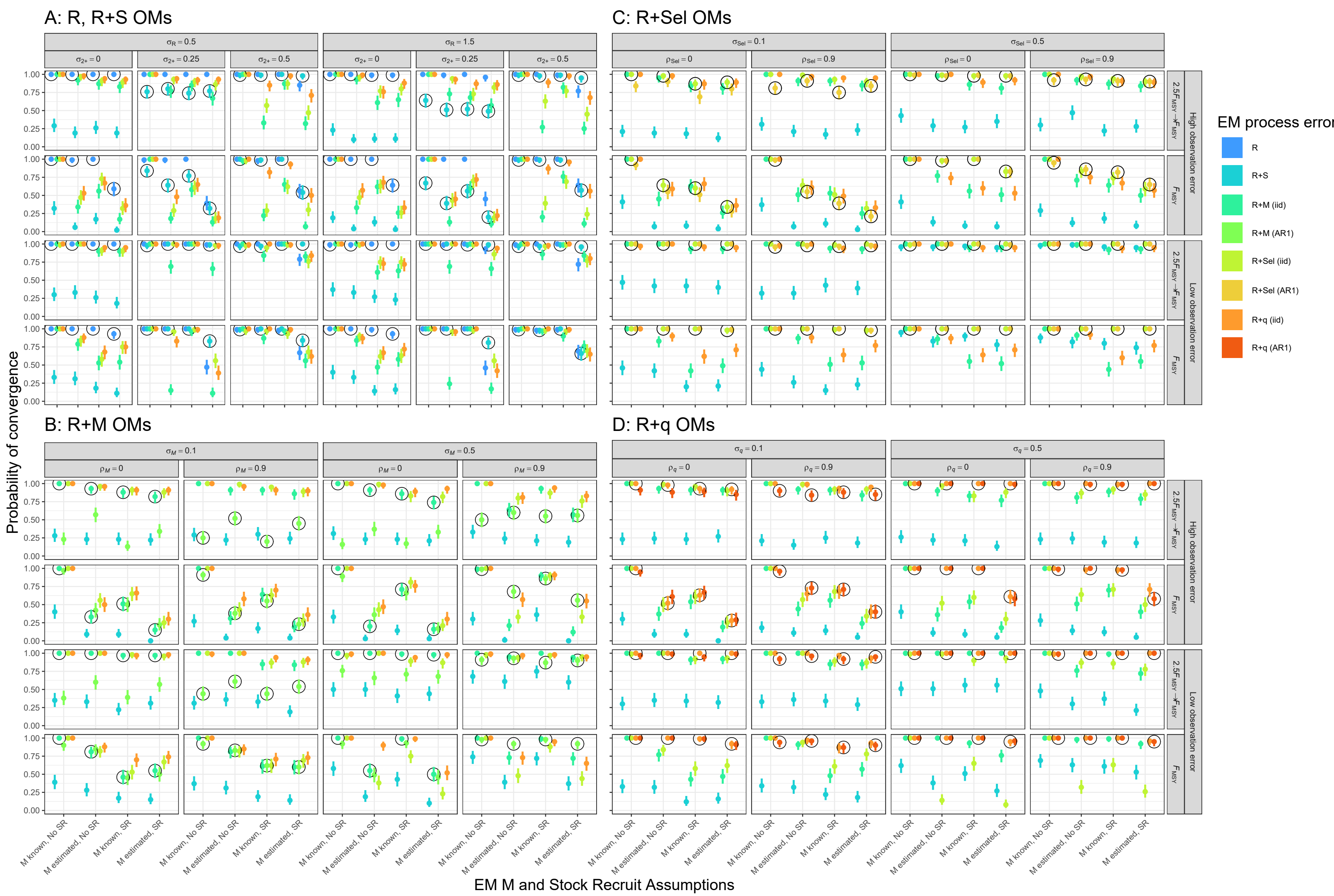

Fig. 1. Estimated probability of fits providing hessian-based standard errors for EMs assuming alternative process error (colored points and lines), and median natural mortality (estimated or known) and Beverton-Holt stock-recruit relationships (estimated or not; along x-axis) when fitted to operating 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.

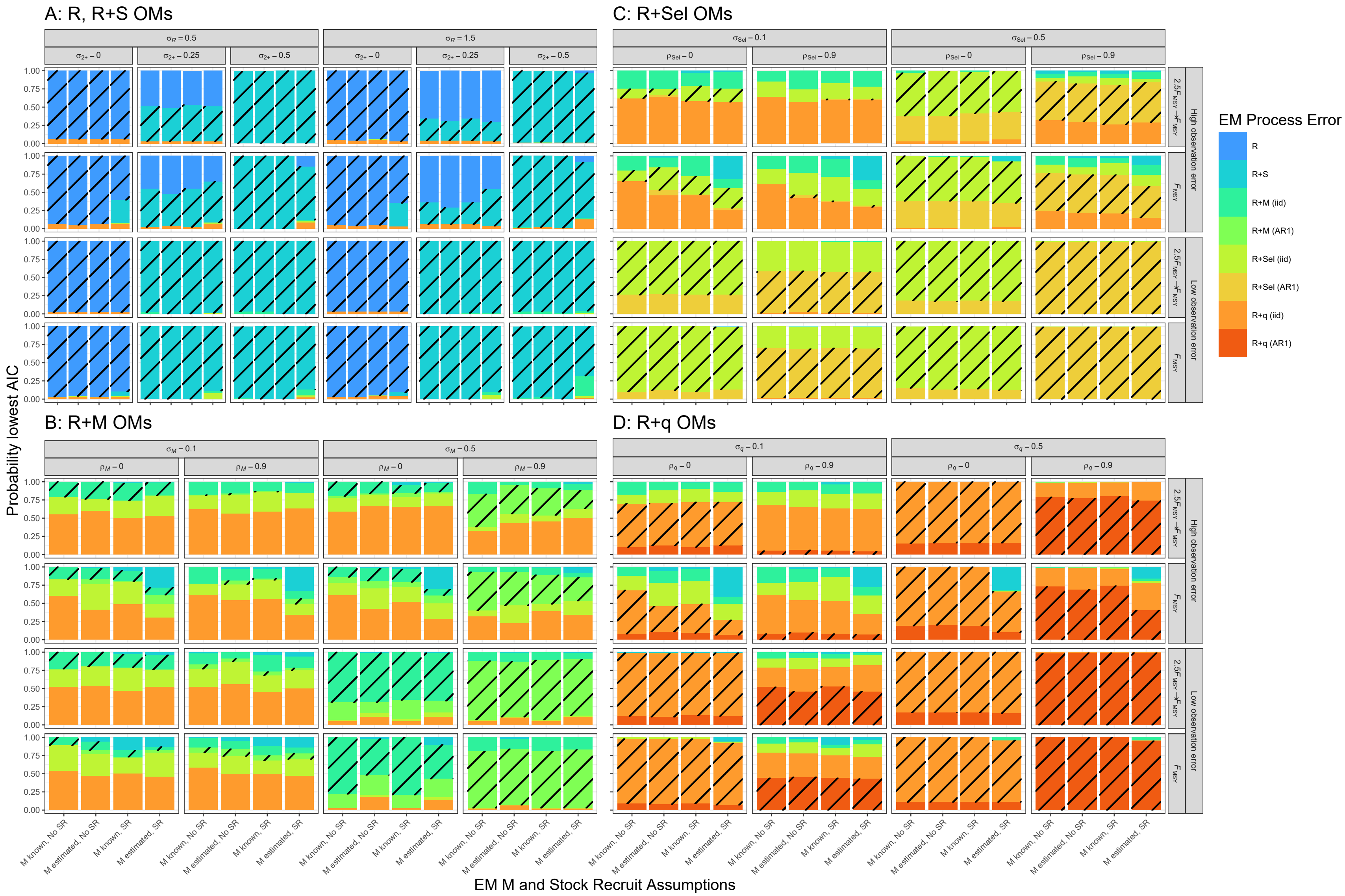

Fig. 2. Estimated probability of lowest AIC for EMs assuming alternative process error structures (colored bars) conditional on alternative assumptions for median natural mortality (estimated or known) and Beverton-Holt stock-recruit relationships (estimated or not; along x-axis) when fitted to operating 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.

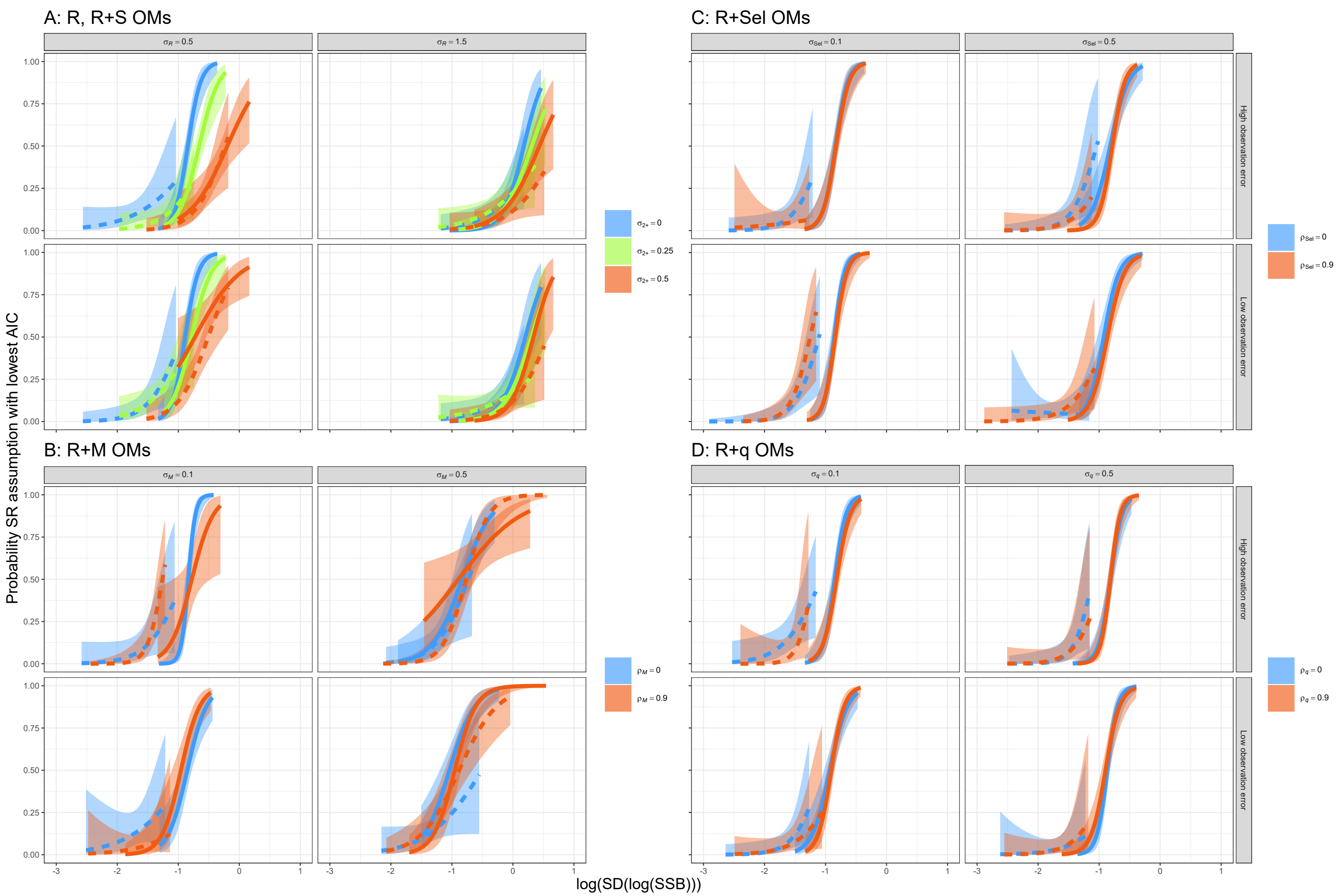

Fig. 3. Estimated probability of lowest AIC from logistic regression on the log-standard deviation of the true log(SSB) in each simulation for estimating model with Beverton-Holt stock-recruit relationships, rather than the otherwise equivalent EM without the stock-recruit relationship. Results are conditional on median M is known in the EM and alternative assumptions EMs having the correct process error structure: R and R+S (A), R+Sel (B), R+M (C), or R+q (D), and median M is assumed known in the EM. Solid and dashed lines are for OMs with and without temporal contrast in fishing pressure, respectively, and polygons represent 95% confidence intervals. Range of results indicates the range of log-standard deviation of log(SSB) for simulations of the particular OM.

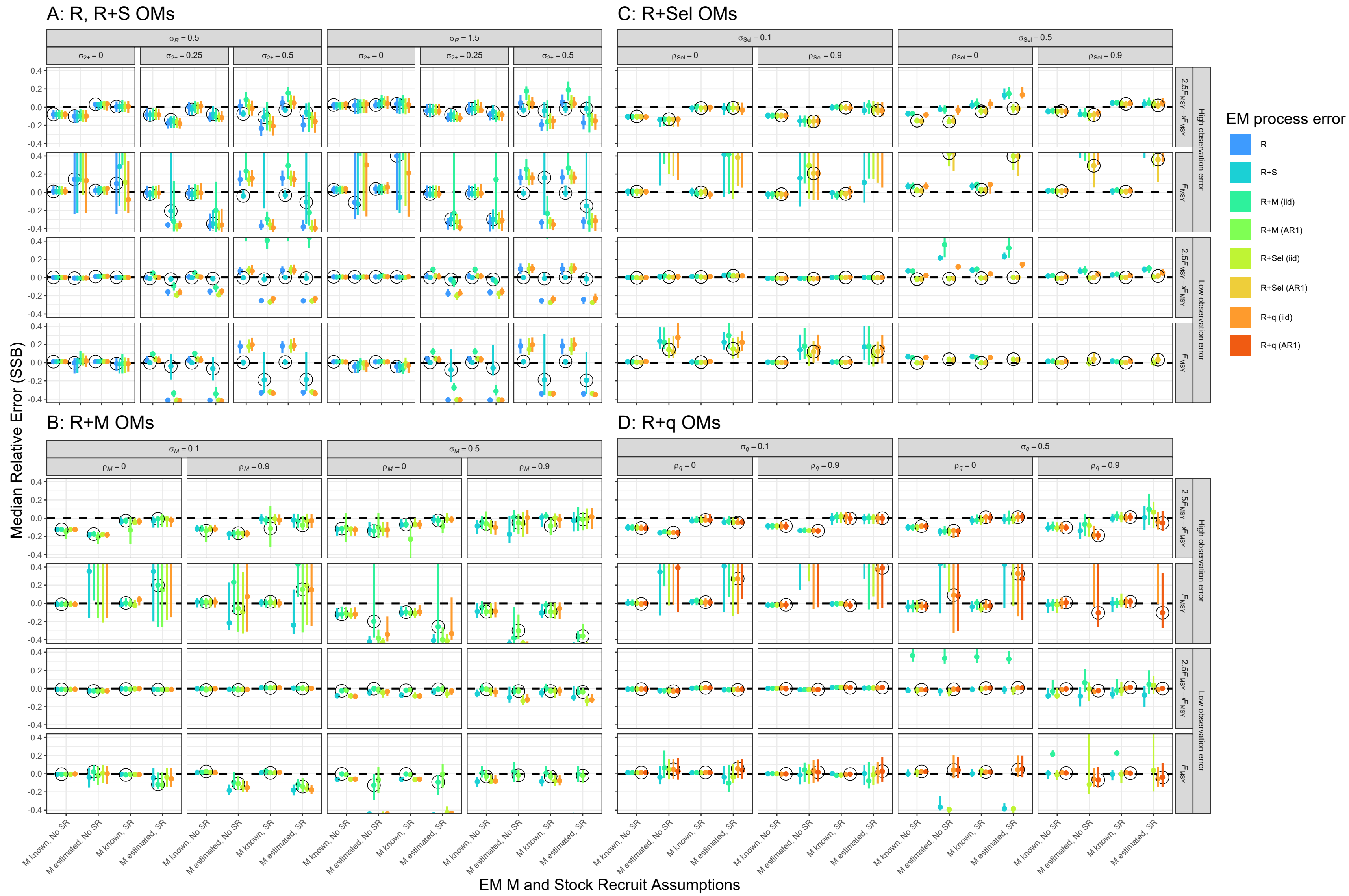

Fig. 4. Median relative error of terminal year SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Diamond shaped points denote results where the EM process error assumption matches that of the operating model.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

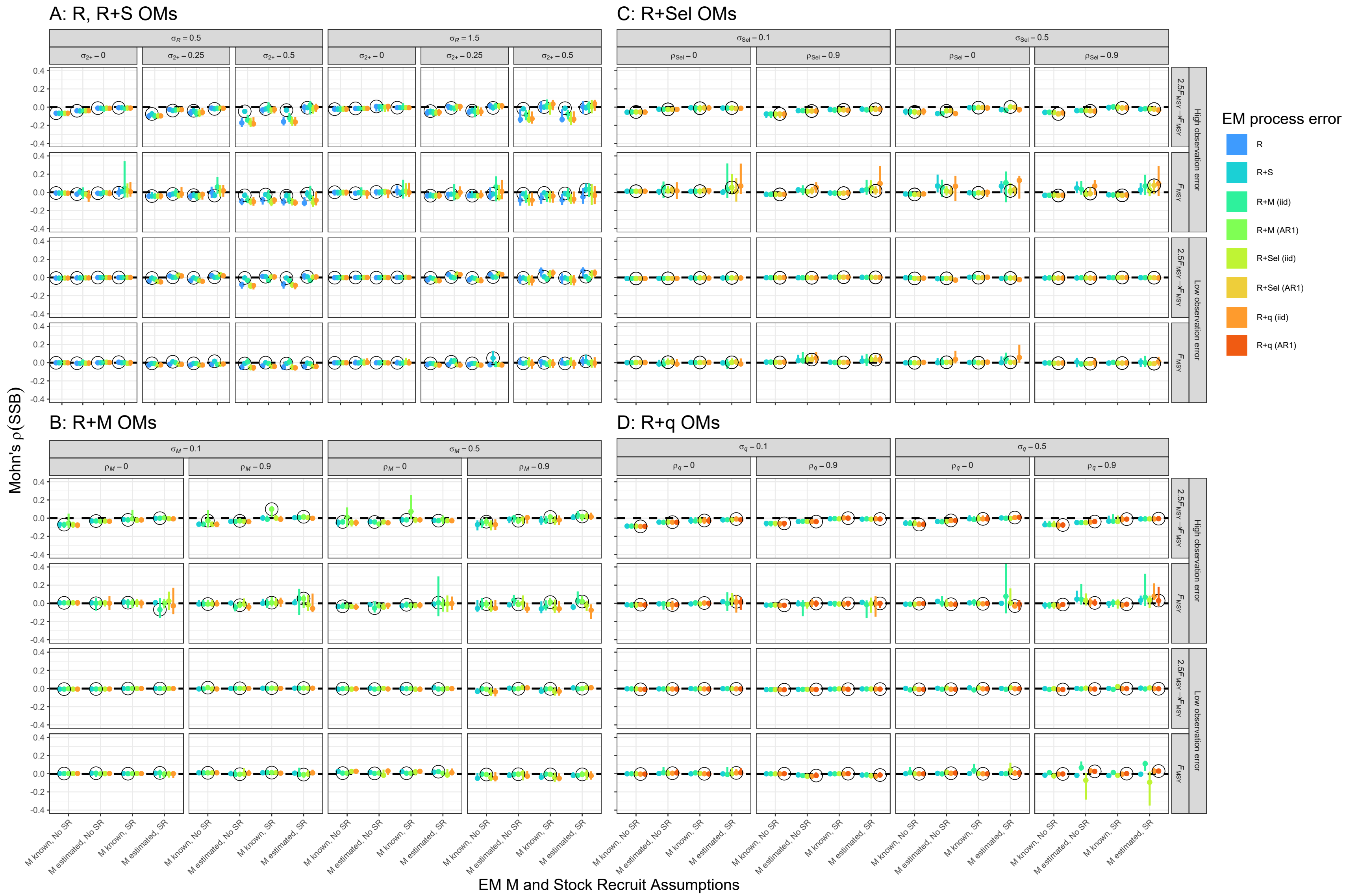

Fig. 5. Median Mohn's rho for SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent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

Supplementary Materials

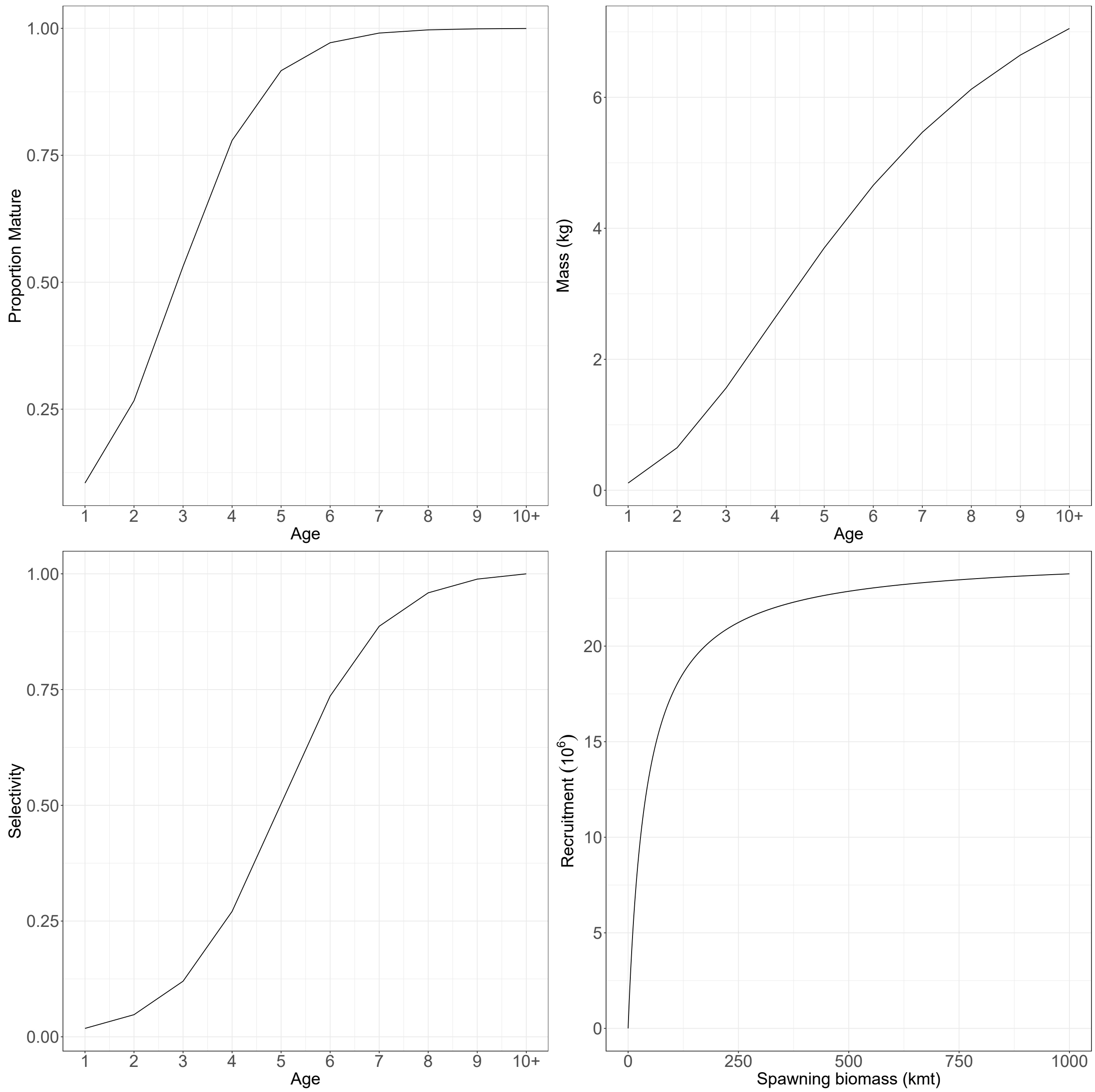

Fig. S1. The proportion mature at age, weight at age, fleet and index selectivity at age, and Beverton-Holt stock-recruit relationship assumed for the population in all operating models. For operating models with random effects on fleet selectivity, this represents the selectivity at the mean of the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

Table S1. Distinguishing characteristics of the operating models with random effects on recruitment and apparent survival (R.R+S). 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	σ_R	σ_{2+}	Fishing History	Observation Uncertainty
NAA_1	0.5		$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_2	1.5		$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_3	0.5	0.25	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_4	1.5	0.25	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_5	0.5	0.50	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_6	1.5	0.50	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_7	0.5		F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_8	1.5		F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_9	0.5	0.25	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_{10}	1.5	0.25	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_{11}	0.5	0.50	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_{12}	1.5	0.50	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
NAA_{13}	0.5		$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{14}	1.5		$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{15}	0.5	0.25	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{16}	1.5	0.25	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{17}	0.5	0.50	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{18}	1.5	0.50	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{19}	0.5		F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
NAA_{20}	1.5		F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
NAA_{21}	0.5	0.25	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
NAA_{22}	1.5	0.25	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
NAA_{23}	0.5	0.50	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
NAA_{24}	1.5	0.50	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5

Table S2. Distinguishing characteristics of the operating models with random effects on recruitment and natural mortality (R+M). 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, σ is defined for the marginal distribution of the processes.

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

Table S3. Distinguishing characteristics of the operating models with random effects on recruitment and selectivity (R+Sel). 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, σ is defined for the marginal distribution of the processes.

Model	σ_R	σ_{Sel}	ρ_{Sel}	Fishing History	Observation Uncertainty
Sel_1	0.5	0.1	0.0	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_2	0.5	0.5	0.0	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_3	0.5	0.1	0.9	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_4	0.5	0.5	0.9	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_5	0.5	0.1	0.0	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
Sel_6	0.5	0.5	0.0	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
Sel_7	0.5	0.1	0.9	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
Sel_8	0.5	0.5	0.9	F_{MSY}	Index SD = 0.1, Age composition SD = 0.3
Sel_9	0.5	0.1	0.0	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{10}	0.5	0.5	0.0	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{11}	0.5	0.1	0.9	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{12}	0.5	0.5	0.9	$2.5F_{\text{MSY}} \to F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{13}	0.5	0.1	0.0	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
Sel_{14}	0.5	0.5	0.0	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
Sel_{15}	0.5	0.1	0.9	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5
Sel_{16}	0.5	0.5	0.9	F_{MSY}	Index SD = 0.4, Age composition SD = 1.5

Table S4. Distinguishing characteristics of the operating models with random effects on recruitment and catchability (R+q). 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, σ is defined for the marginal distribution of the processes.

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

Table S5. Distinguishing characteristics of the estimating models and operating model process error categories (R, R+S, R+M, R+Sel, R+q) where used.

Model	Recruitment model	Median M	Process error	R,R+S OMs	R+M OMs	R+Sel OMs	R+q OMs
EM_1	Mean recruitment	0.2	R ($\sigma_{2+} = 0$)	+	−	−	−
EM_2	Beverton-Holt	0.2	R ($\sigma_{2+} = 0$)	+	−	−	−
EM_3	Mean recruitment	Estimated	R ($\sigma_{2+} = 0$)	+	−	−	−
EM_4	Beverton-Holt	Estimated	R ($\sigma_{2+} = 0$)	+	−	−	−
EM_5	Mean recruitment	0.2	R+S (σ_{2+} estimated)	+	+	+	+
EM_6	Beverton-Holt	0.2	R+S (σ_{2+} estimated)	+	+	+	+
EM_7	Mean recruitment	Estimated	R+S (σ_{2+} estimated)	+	+	+	+
EM_8	Beverton-Holt	Estimated	R+S (σ_{2+} estimated)	+	+	+	+
EM_9	Mean recruitment	0.2	R+M ($\rho_M = 0$)	+	+	+	+
EM_{10}	Beverton-Holt	0.2	R+M ($\rho_M = 0$)	+	+	+	+
EM_{11}	Mean recruitment	Estimated	R+M ($\rho_M = 0$)	+	+	+	+
EM_{12}	Beverton-Holt	Estimated	R+M ($\rho_M = 0$)	+	+	+	+
EM_{13}	Mean recruitment	0.2	R+Sel ($\rho_{\mathrm{Sel}} = 0$)	+	+	+	+
EM_{14}	Beverton-Holt	0.2	R+Sel ($\rho_{\mathrm{Sel}} = 0$)	+	+	+	+
EM_{15}	Mean recruitment	Estimated	R+Sel ($\rho_{\mathrm{Sel}} = 0$)	+	+	+	+
EM_{16}	Beverton-Holt	Estimated	R+Sel ($\rho_{\mathrm{Sel}} = 0$)	+	+	+	+
EM_{17}	Mean recruitment	0.2	R+q ($\rho_q = 0$)	+	+	+	+
EM_{18}	Beverton-Holt	0.2	R+q ($\rho_q = 0$)	+	+	+	+
EM_{19}	Mean recruitment	Estimated	R+q ($\rho_q = 0$)	+	+	+	+
EM_{20}	Beverton-Holt	Estimated	R+q ($\rho_q = 0$)	+	+	+	+
EM_{21}	Mean recruitment	0.2	R+M (ρ_M estimated)	−	+	−	−
EM_{22}	Beverton-Holt	0.2	R+M (ρ_M estimated)	−	+	−	−
EM_{23}	Mean recruitment	Estimated	R+M (ρ_M estimated)	−	+	−	−
EM_{24}	Beverton-Holt	Estimated	R+M (ρ_M estimated)	−	+	−	−
EM_{25}	Mean recruitment	0.2	R+Sel (ρ_{Sel} estimated)	−	−	+	−
EM_{26}	Beverton-Holt	0.2	R+Sel (ρ_{Sel} estimated)	−	−	+	−
EM_{27}	Mean recruitment	Estimated	R+Sel (ρ_{Sel} estimated)	−	−	+	−
EM_{28}	Beverton-Holt	Estimated	R+Sel (ρ_{Sel} estimated)	−	−	+	−
EM_{29}	Mean recruitment	0.2	R+q (ρ_q estimated)	−	−	−	+
EM_{30}	Beverton-Holt	0.2	R+q (ρ_q estimated)	−	−	−	+
EM_{31}	Mean recruitment	Estimated	R+q (ρ_q estimated)	−	−	−	+
EM_{32}	Beverton-Holt	Estimated	R+q (ρ_q estimated)	−	−	−	+

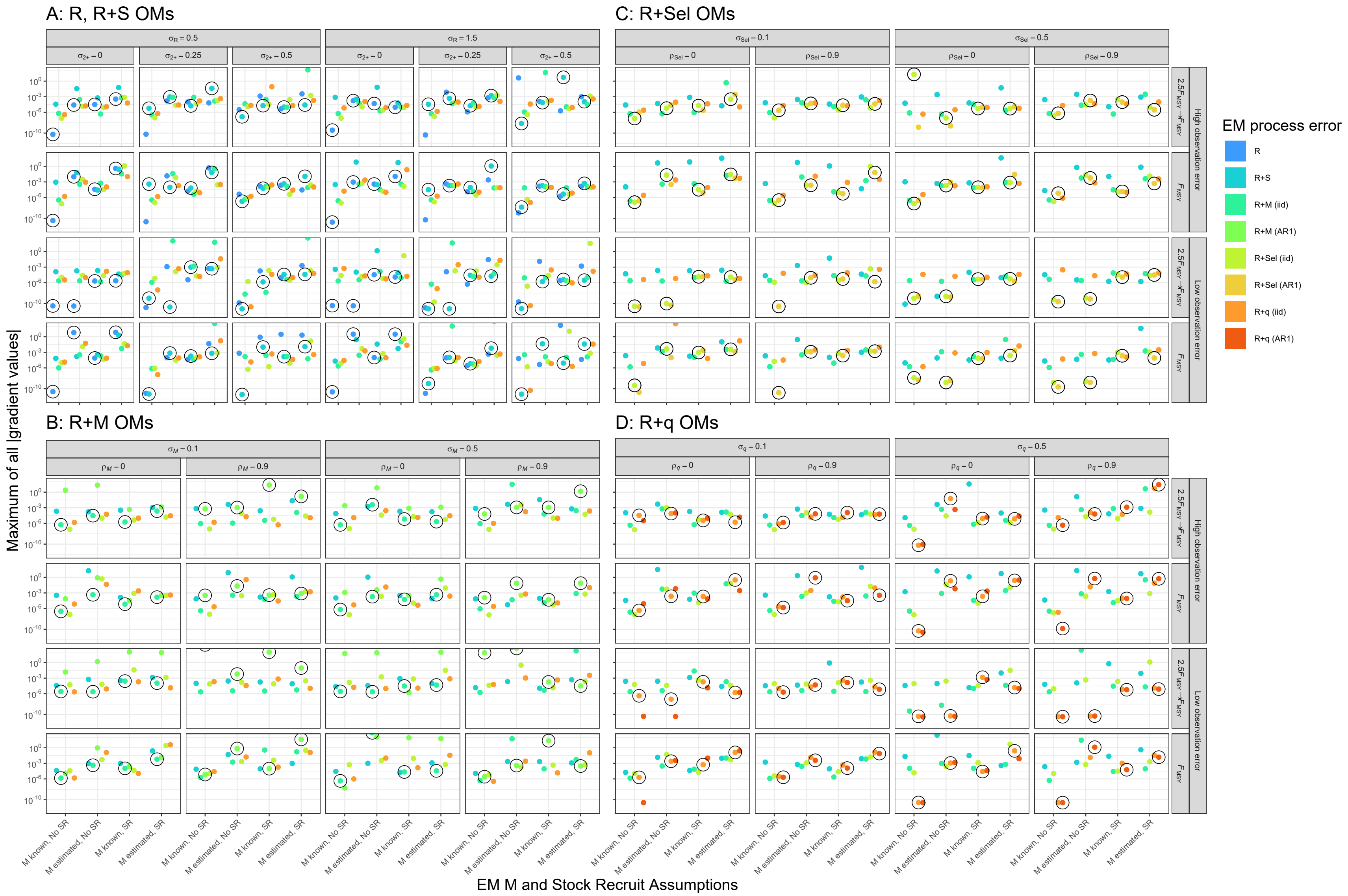

Fig. S2. The maximum of the absolute values of all gradient values for all fits that provided hessian-based standard errors across all simuated data sets of a given OM configuration (A: R and R+S, B: R+M, C: R+Sel, or D: R+q). Results are conditional on EM fits with alternative process error type (colored points and lines), median natural mortality (estimated or known) and recruitment assumptions (Beverton-Holt stock-recruit relationship 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.

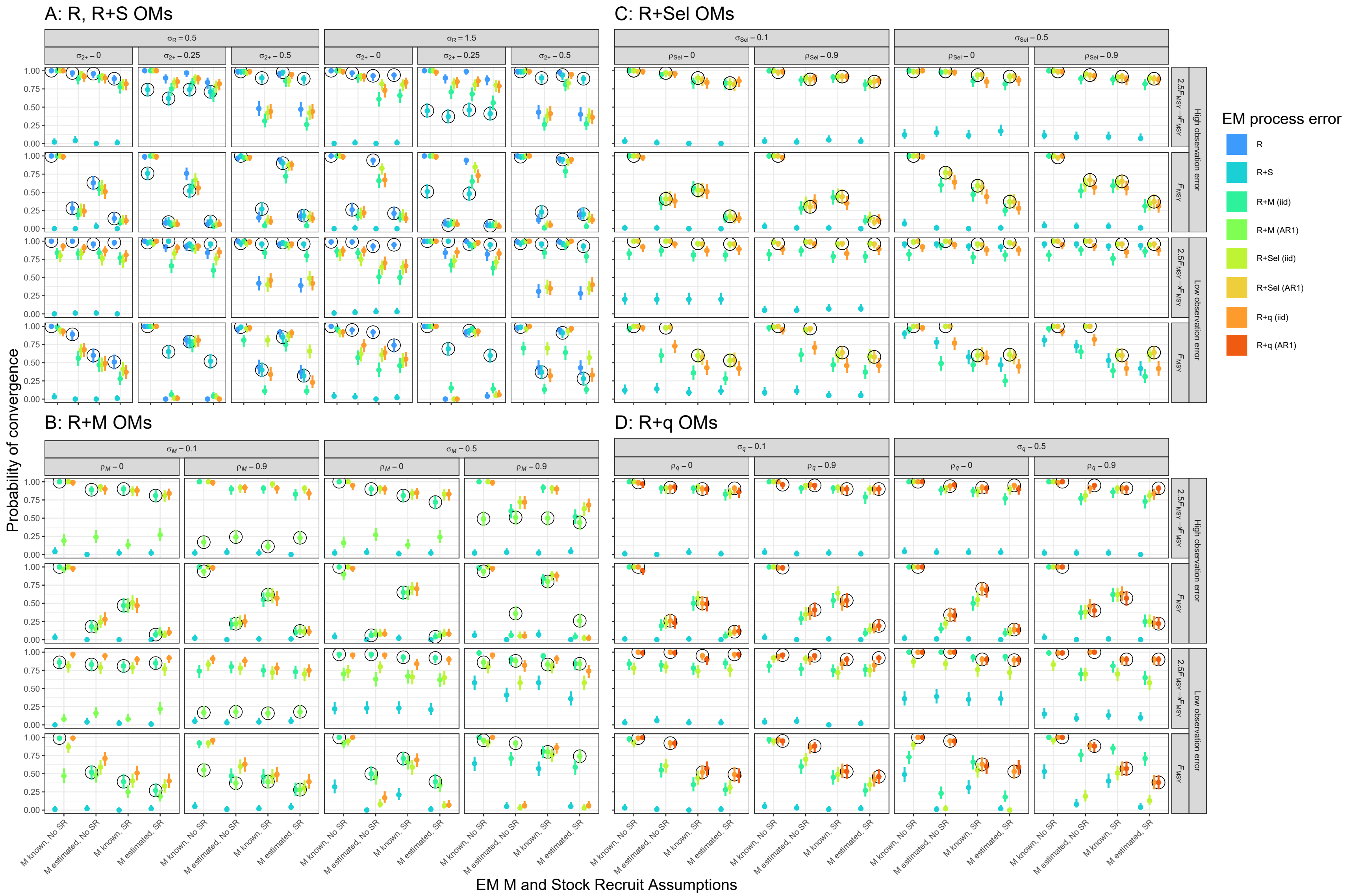

Fig. S3. Probability of estimating models providing maximum absolute values of gradients less than 10^{-6} assuming alternative process error (colored points and lines), and median natural mortality (estimated or known) and Beverton-Holt stock-recruit relationships (estimated or not; along x-axis) when 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.

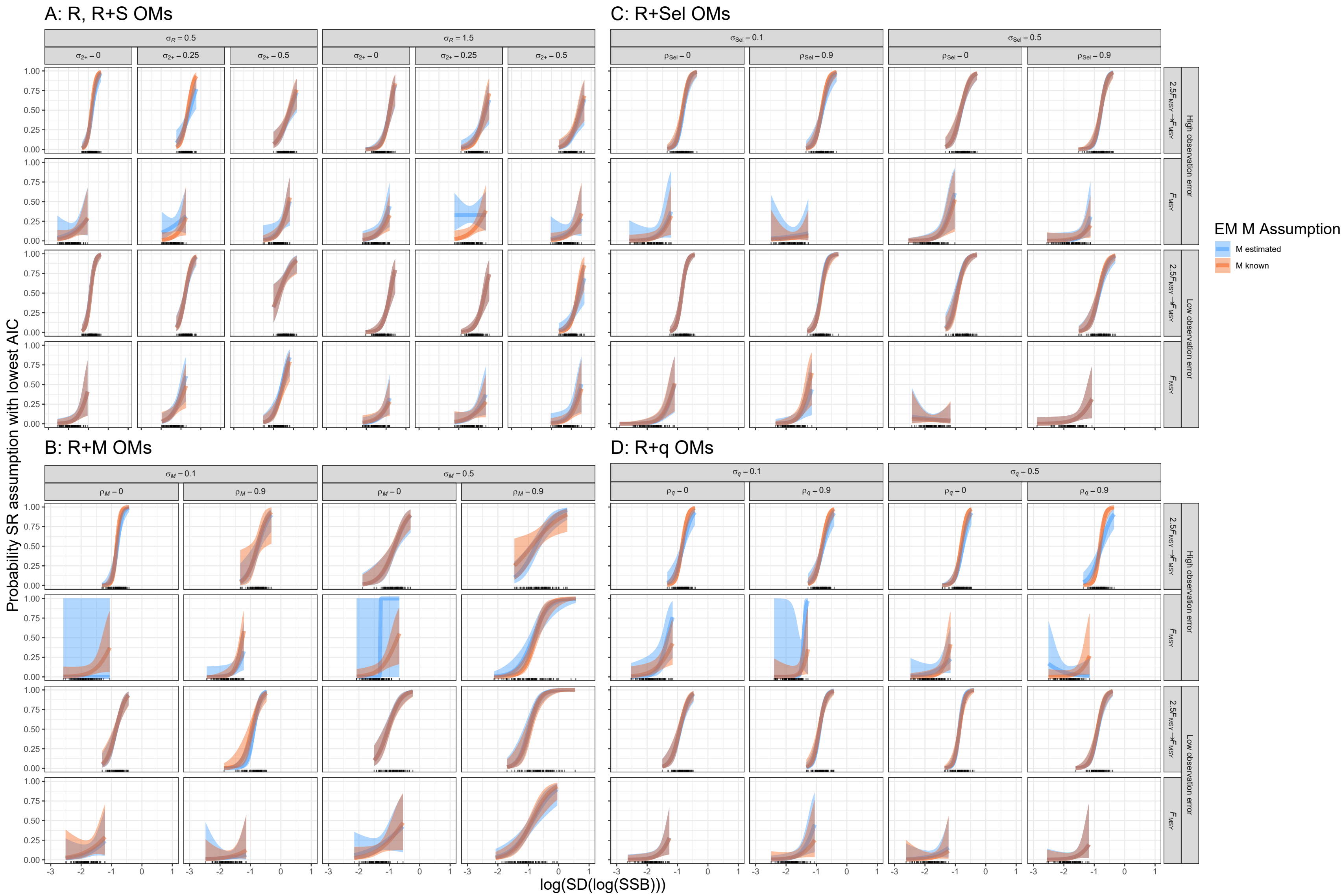

Fig. S4. Estimated probability of lowest AIC from logistic regression on the log-standard deviation of the true log(SSB) in each simulation for estimating model with Beverton-Holt stock-recruit relationships, rather than the otherwise equivalent EM without the stock-recruit relationship. Results are 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.

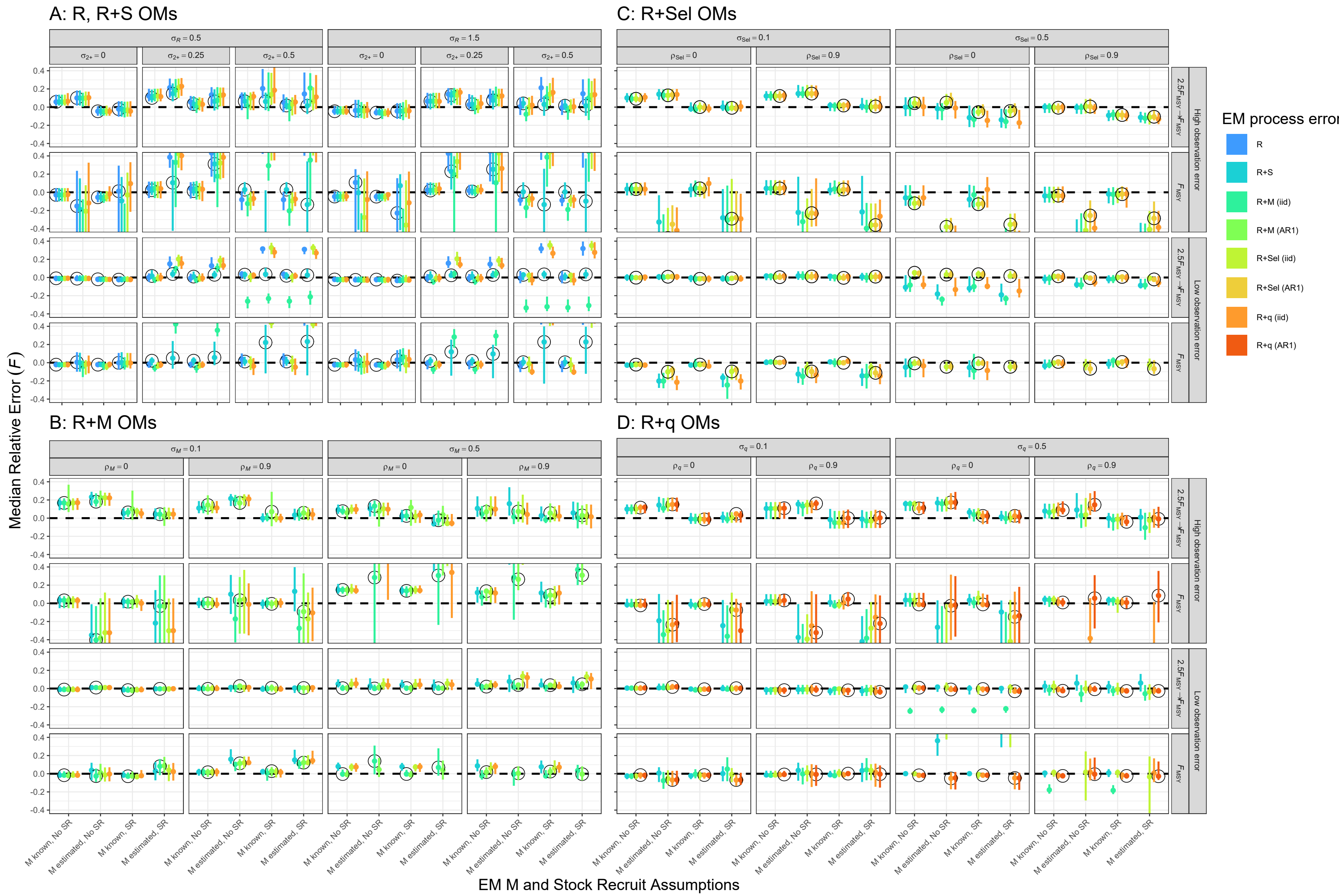

Fig. S5. Median relative error of terminal year fully-selected fishing mortality for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that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

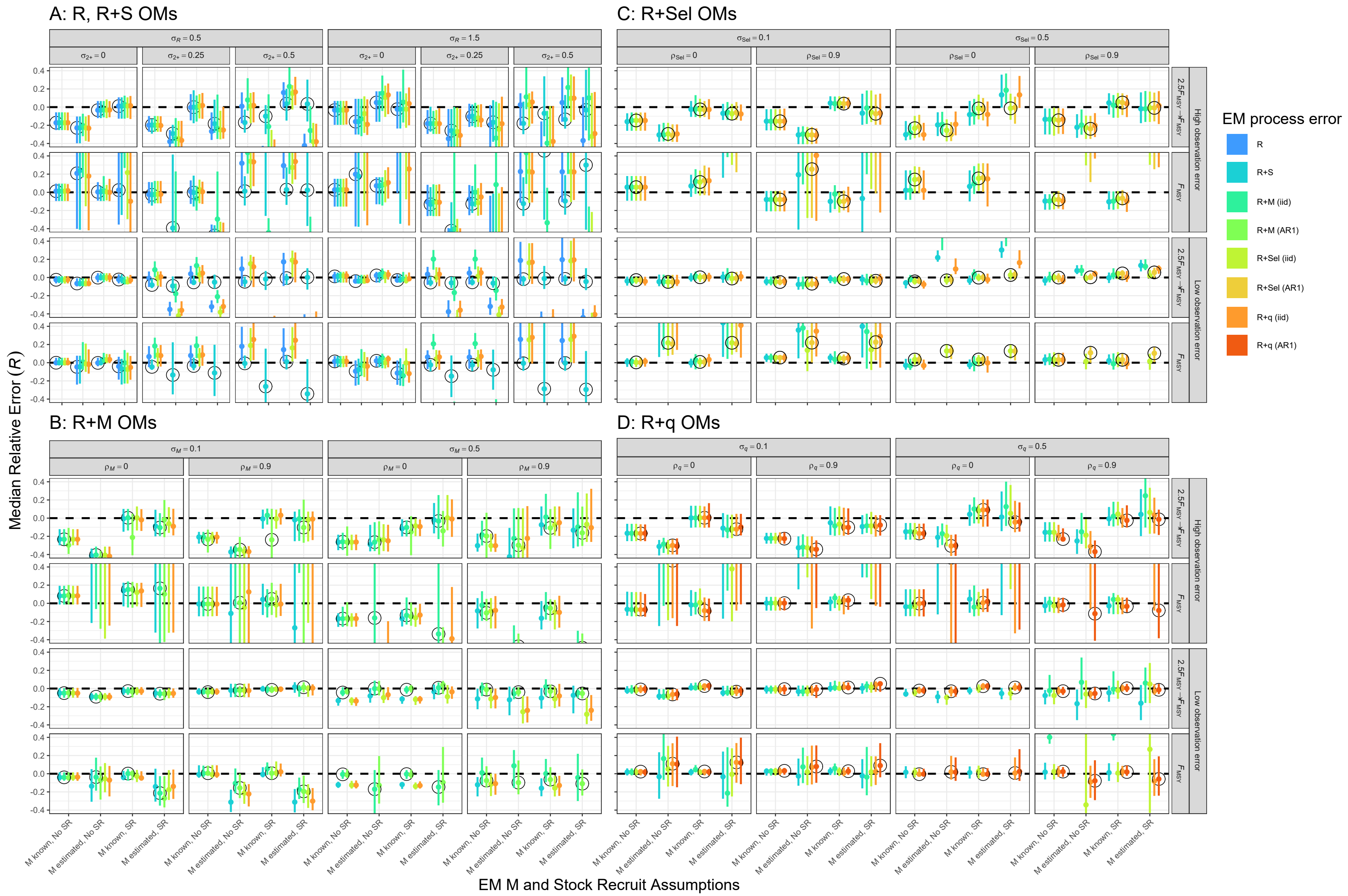

Fig. S6. Median relative error of terminal year recruitment for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating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

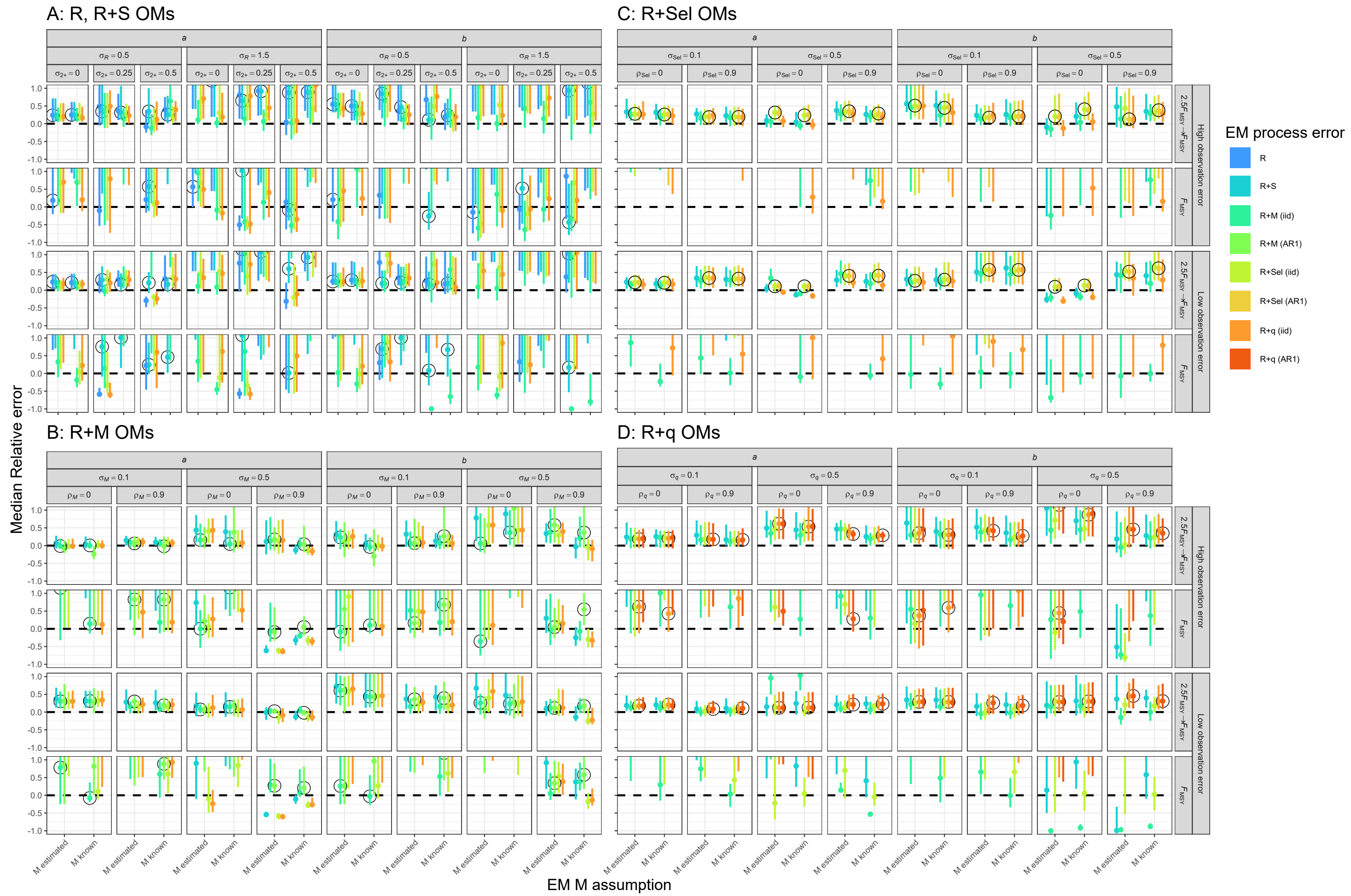

Fig. S7. Median relative error of Beverton-Holt stock-recruit parameters (a and b) for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches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

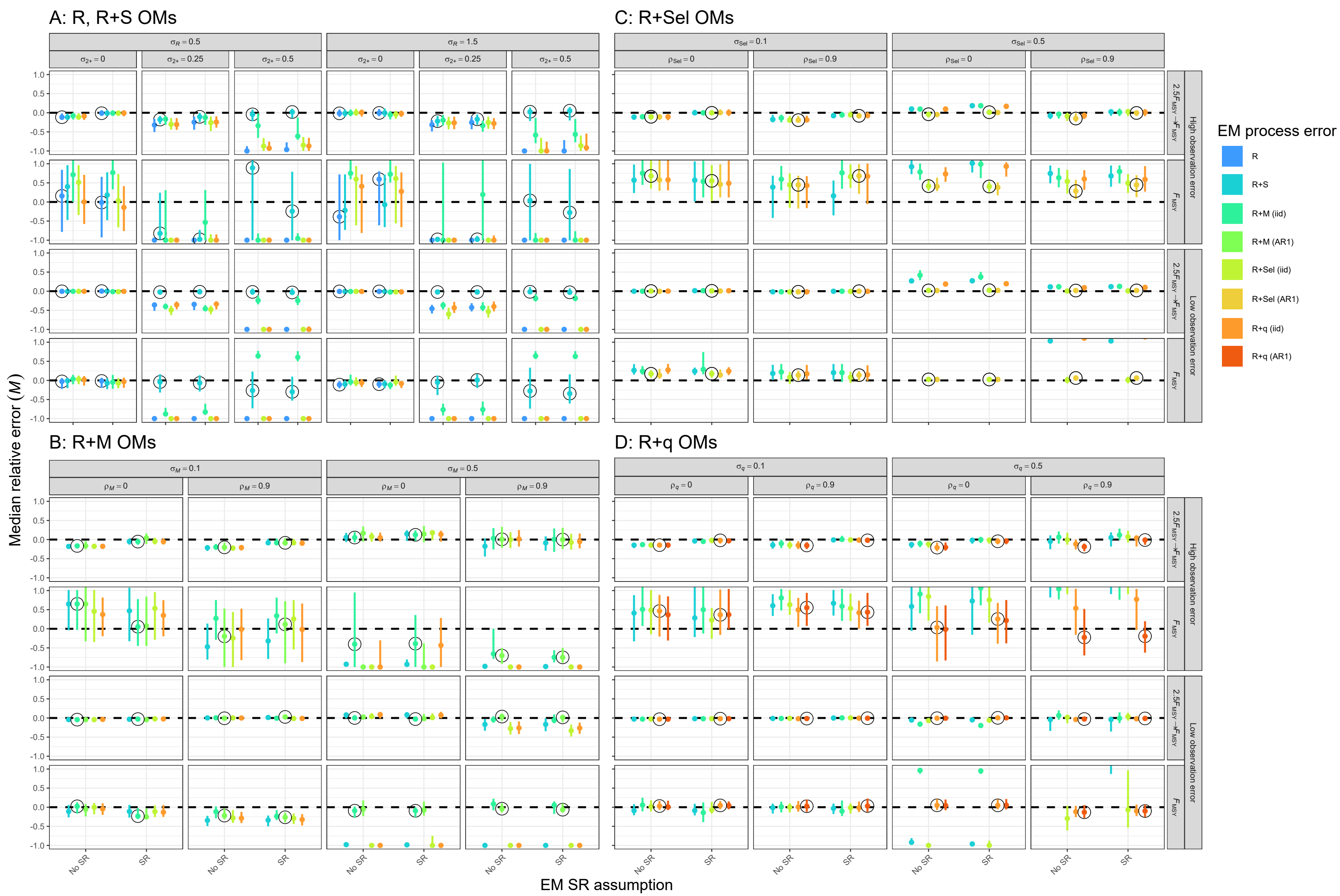

Fig. S8. Median relative error of median natural mortality for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

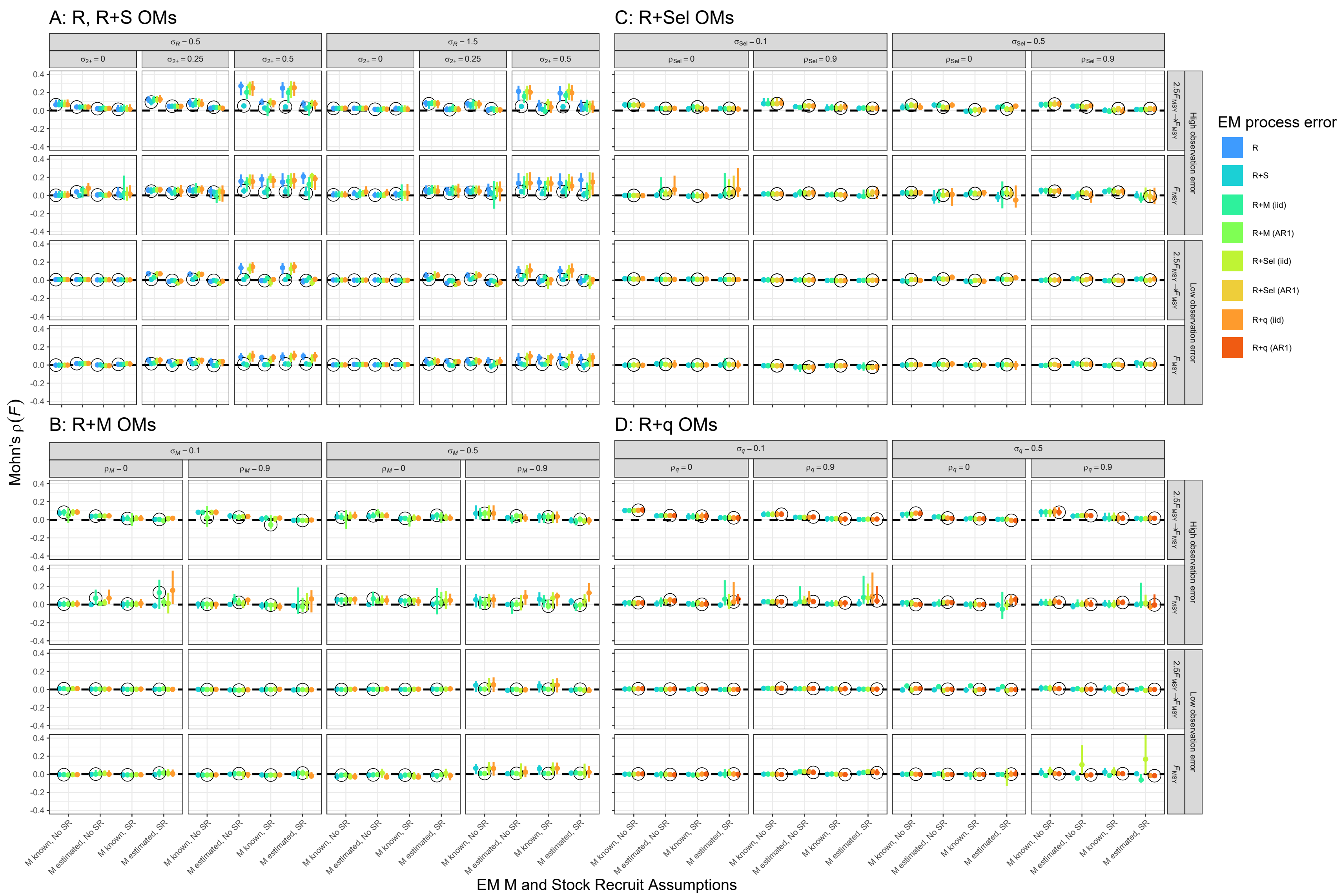

Fig. S9. Median Mohn's ρ of fishing mortality averaged over all age classes for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of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

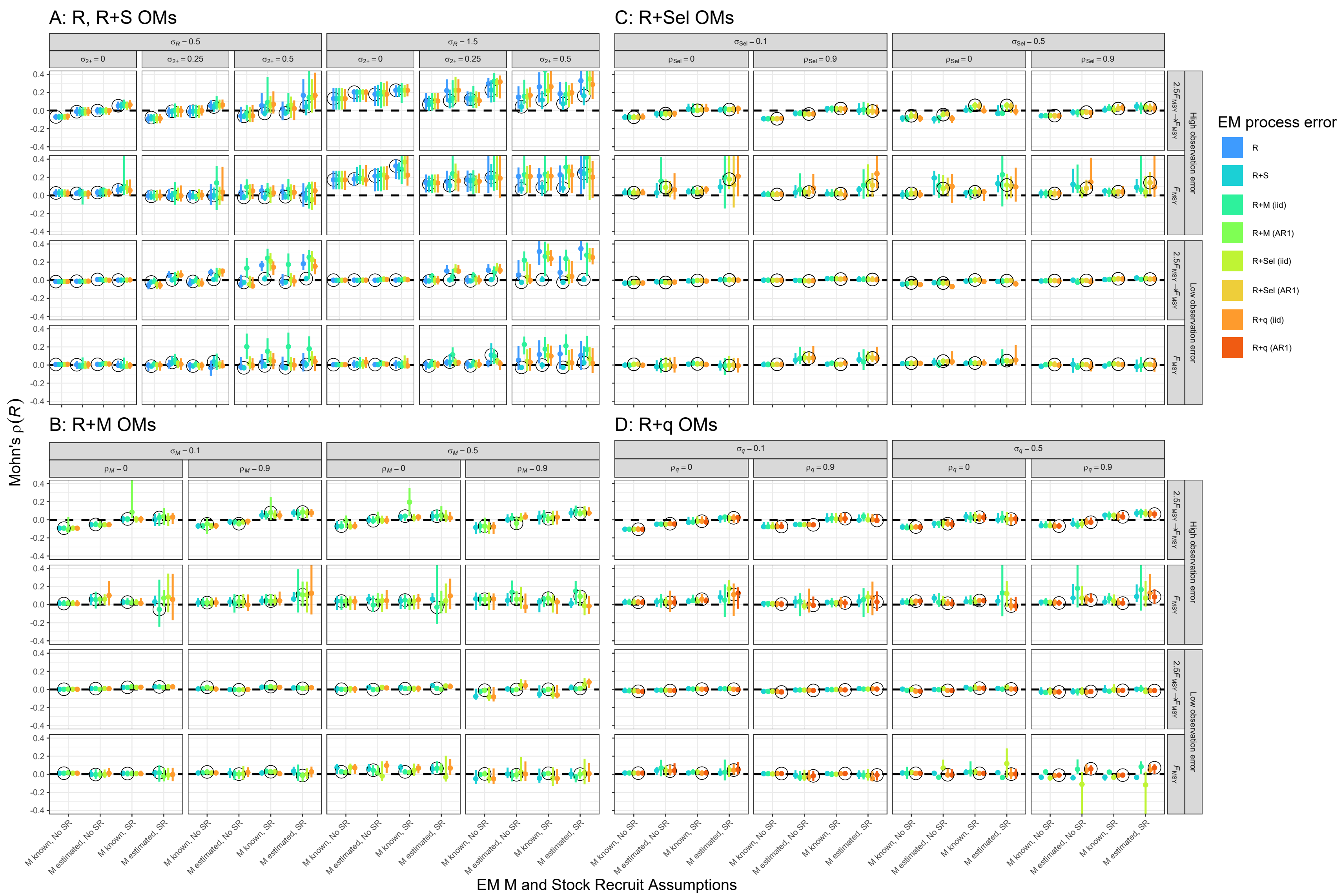

Fig. S10. Median Mohn's ρ of recruitment for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.