An investigation of factors affecting inferences from and reliability 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]

~~02 May~~29 September, 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, and other process error sources when variability was greater. 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; Pontavice et al. 2022), but 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). Finally, although the importance of autocorrelation in process errors is recognized, investigations of the ability to distinguish state-space assessment models with and without autocorrelation and whether such misspecification is detrimental to estimation of important population metrics are lacking (Johnson et al. 2016; Xu et al. 2019).

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 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 NEUS fish stocks (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 Northeast Unite States (NEUS) assessments using WHAM.

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). 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 biological inputs to calculations of spawning biomass per recruit (i.e., weight, maturity, and natural mortality at age) are constant in the apparent survival (R+S) selectivity (R+Sel), and survey catchability (R+q) 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). We assumed a value of 0.2 for the natural mortality rate in OMs without process errors on natural mortaliyt and for the median rate for OMs with process errors on natural mortality.

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 S3).

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 $\epsilon_{M,y} \sim \mathrm{N}\left(0, \sigma_M^2\right)$. The marginal standard deviations we assumed for log natural mortality random effects were $\sigma_M \in \{0.1, 0.5\}$ and the random effects were either uncorrelated or first-order autoregressive (AR1, $\rho_M \in \{0, 0.9\}$). Uncorrelated random effects were also included on recruitment with $\sigma_R = 0.5$ (hence, we denote these OMs as R+M). 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 S4).

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 S5).

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 S6).

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 S7). 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 S7 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. As such, EMs would either be configured completely correctly for the OM, or there could be mis-specification in assumptions of process error autocorrelation, the type of process error, or the SRR (Beverton-Holt or none).

Measures of reliability

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~~ probability of convergence as measured by occurrence of a positive-definite hessian matrix at the optimized negative log-likelihood that could be inverted (i.e., providing hessian-based standard error estimates). We also provide results in the Supplementary Materials for ~~the maximum~~ convergence defined by the maximum absolute gradient $< 1^{-6}$ and the maximum of the absolute ~~values among all gradients~~ gradient values 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~~ also investigated the reliability of AIC-based model selection for two purposes. First, we analyzed selection of each process error model structure (R, R+S, R+M, R+Sel, R+q) using marginal AIC. For a given ~~operating model~~OM simulated data set, we compared AIC for EMs ~~that were all configured the same for~~ with different process error assumption conditional on whether median natural mortality ~~(known or estimated) and the SRR~~ (rate and the Beverton-Holt ~~or none).~~ ~~We also estimated the probability of correctly selecting~~ stock-recruit relationship were estimated. We tabulated the models providing the lowest AIC across simulated data sets. Second, we analyzed AIC-based selection between EMs with and without the Beverton-Holt ~~SRR assumed and models without the SRR (null model). We made these comparisons between models that otherwise assumedthe same process error structure as the operating model and both of the compared models either estimate median natural mortality or assume it is known~~stock-recruit relationship assumed. 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~~ indicated generally poor performance of AIC in determining the Beverton-Holt model 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~~ also considered the effect of of the log-standard deviation of the true log(SSB) (similar to the log of the coefficient of variation for SSB) on model selection since simulations with realized SSB producing low and high ~~recruitments~~

recruitment would have larger variation in realized SSB.

All model selection results condition ~~on whether all of the compared estimating models completed the~~ only on completion of the optimization process without failure for all of the compared EMs. We did not condition on convergence as defined ~~by a gradient threshold or invertibity of the hessian~~ above 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

We also investigated bias in estimation of various model attributes as a measure of reliability. For a given model attribute we calculated the relative error

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

$$\mathrm{RE}\left(\theta_i\right) = \frac{\widehat{\theta}_i - \theta_i}{\theta_i} \tag{1}$$

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~~ analyzed simulation results for estimates of 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 were strongly negatively correlated with those for SSB and are provided in the Supplementary Materials.~~

Mohn's ρ

Finally, we investigated presence of retrospective patterns in fitted models as a measure of reliability. We calculated Mohn's ρ for SSB, ~~fully-selected fishing mortality~~ fishing mortality (averaged over all age classes), and recruitment for each EM fit to each OM simulated data set (Mohn 1999). We fit ~~7 peels for each EM and calculated median 95% confidence intervals for~~ $P = 7$ peels to each simulated data set and calculated Mohn's ρ ~~using the same methods as that for relative bias~~for a given attribute θ as

$$\rho\left(\theta\right) = \frac{1}{P}\sum_{p=1}^{P}\frac{\widehat{\theta}_{Y-p,Y-p} - \widehat{\theta}_{Y-p,Y}}{\widehat{\theta}_{Y-p,Y}} \tag{2}$$

where $\widehat{\theta}_{i,j}$ is the estimate for attribute θ in year i from a model fit using data up to year j.

~~Results~~

Summarizing results across OM and EM attributes

~~For many R and R+S OMs, convergence rate declined when either the median natural mortality rate or the Beverton-Holt SRR was estimated even when the process error assumptions of the EMs and OMs matched (Figure S14, 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~~ The measures of central tendency and variability of observed values for specific OM and EM attributes (e.g., low or high observation error) that we described above can indicate scenarios that provide better or poorer reliability. However, the OM and EM attributes that we investigated are numerous, so we used two methods to summarize the most important factors for differences in results. The first method was fitting regression models with the response being each of the measures of reliability described above and predictor variables were defined based on OM and EM characteristics (e.g., MacKinnon et al. 1995; Wang et al. 2017; Harwell et al. 2018). For the binary indicators of convergence and AIC-based selection of a stock-recruit relationship, we performed logistic regressions. For indicators of AIC-based selection of 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,~~ type (multiple categories) we performed multinomial regressions. For other measures of reliability we fit linear regression models to transformed responses. Because relative errors (Eq. 1) 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.~~ Mohn's ρ for the various parameters are bounded below at -1, we used a transformation of these values

$$y_i = \log\left[f\left(\widehat{\theta}_i, \theta_i\right) + 1\right] \tag{3}$$

where f is either the relative error (Eq. 1) or Mohn's ρ (Eq. 2) for simulation i, so that values are unbounded. For relative errors, y_i is the log-scale error. We omitted simulation estimates equal to zero (RE = -1). For all regressions we fit separate models with individual factors included, with all factors combined, with including all second order interactions, and including all third order interactions. For the multinomial regression, we used the `vglm` function from the VGAM package (Yee 2008; Yee 2015). We tabulated percent reduction in residual deviance for each of regression fits. We did not perform formal statistical analyses of effects of OM and EM attributes on results (e.g., ANOVA) because of the lack of independence of the "observations" that results from fitting multiple EMs to each simulated data set.

~~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 S14, B)~~ The second method involved fitting classification and regression trees (Breiman et al. 1984) to show how the OM and EM attributes, and their interactions, split the values for each measure of reliability (e.g., Gonzalez et al. 2018; Collier et al. 2022). We used classification trees for categorical measures (convergence and AIC) and regression trees for the other measures with continuous scales (relative error and Mohn's ρ). The response variables were the same as the regressions for the deviance reduction analyses. We used the `rpart` function in the rpart package (https://cran.r-project.org/package=rpart) to fit trees. Full trees were determined using default settings except that we increased the number of cross-validations to 100. For clarity, we pruned the full trees to show just the primary branches.

We also provide detailed results for all measures of reliability at each combination of OM and EM attributes in the Supplementary Materials. For confidence intervals of probability of convergence, we used the Clopper-Pearson exact method (Clopper and Pearson 1934; Thulin 2014). For AIC selection of process error configuration we provide estimates of the proportions of simulations where each EM type was selected. For AIC selection of the stock-recruit relationship, we provide predicted probabilities from logistic regressions as a function of SSB variability for each OM and EM type. 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, and Mohn's ρ using the binomial distribution approach as in Miller and Hyun (2018) and Stock and Miller (2021). For each EM we calculated median and 95% confidence intervals using the same methods as that for relative bias.

Results

Convergence performance

For probability of convergence, EM process error assumption was the single attribute that resulted in the largest percent reduction in deviance for all OM process error types other than R+~~M EMs that estimated autocorrelation of process errors had poor convergencefor 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 S14, C).~~ OMs where the EM M assumption explained the most residual deviance (Table 1). However, including interactions of OM and EM factors also provided large reductions in residual deviance, suggesting successful convergence depended on a combination OM and EM attributes. Classification trees for each OM process error type, all had the primary branch defined using the same attribute that provided the largest reduction in deviance (Figure 1). EMs that assumed R+~~Sel EMs generally converged better than other EMs for~~ S process errors converged poorly for all OMs that were simulated with the alternative process error assumptions (R, 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~~ M, R+~~q OMs, convergence of~~ Sel, an R+q ~~EMs was generally better than that of other EMs when there was contrast~~ OMs). For all trees, branches based on the OM fishing mortality history showed better convergence when the OM included a change in fishing pressure~~(Figure (S14, D). Convergence of R+S EMs was generally worse than that of all~~

~~other EMs across all OMs whether or not~~. Branches based on whether the Beverton-Holt stock-recruit model was assumed or not, showed better convergence when it was not estimated and branches based on whether the EM estimated the 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~~rate showed better convergence when it was treated as known. For certain R+M and R+Sel OMs, better convergence was also observed when there was lower observation uncertainty.

When convergence is defined by a gradient threshold, the primary factor explaining deviance reduction is the same for all OM process error types, but there are some differences in deviance reduction for secondary factors (Table S8), and probability of convergence, overall, was lower (Figure S2). We found a wide range of maximum absolute values of gradients for models that ~~converged~~ had invertible hessians (Figure S3). 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

~~Marginal AIC accurately determined~~

Process error structure

For AIC selection of the correct process error ~~assumptions in EMs when data were generated from R and~~ configuration, the magnitude of observation and process error were the attributes that resulted in the largest percent reductions in deviance across OM process error types other than R OMs (Table 2). Both variance of apparent survival random effects (σ_{2+}) and degree of observation error explained the largest reductions for R+S ~~OMs~~and R+Sel OMs, whereas variance of process errors provided the largest reductions in R+M and R+q OMs. Comparatively, none of the OM or EM attributes explained particularly large reductions in deviance for R OMs, but fishing history, whether a stock-recruit model was estimated, ~~regardless of whether median natural mortality or a SRR was estimated(Figure S16, A). Attempting to estimate~~ and whether 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.~~ was known or estimated provided similar and the largest reductions. Inclusion of second and third order interactions, did not provide large reductions in deviance for any of the OM process error types.

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 S16, 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 OMswith 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~~ all OM process error types other than R OMs, the attributes defining the primary branches of classification trees matched those that provided the largest reductions in deviance (Figure 2). Across all OMs, AIC was more accurate for the process error type when process error variability was greater and when observation error was lower. No branches were estimated for classification trees fit to the R OMs, likely because accuracy was high across all simulations (0.94), although inspection of the fine-scale results shows there is some degradation in AIC selection when a stock recruit relationship and median natural mortality rate are estimated for R OMs with constant fishing pressure and high observation error (~~Figures S16, C). When there was low variability in selectivity process errors and high observation error,~~ Figure S16, top left).

Stock-recruit relationship

Logistic regressions for AIC selection of the Beverton-Holt stock recruit relationship, showed OM fishing history and variation in SSB ($\log \mathrm{SD}_{\mathrm{SSB}}$) provided substantial reductions in deviance for R+~~q or~~ M, 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~~ Sel, and R+q OMs ~~with high variability in catchability process errors (Figures S16,D). The R~~ (Table 3). For R OMs, fishing history provided the largest reduction in deviance whereas none of the attributes individually provided large reductions in deviance for R+~~q OMswith 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.~~

~~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~~ S 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 S17).~~

~~With~~ inclusion of all attributes provided larger reductions in deviance than the sum of individual contributions for both 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 S17, A). The largest~~ OMs. For all OM process error types, inclusion of interaction terms provided relatively little reduction in residual deviance.

Attributes defining the primary branches of classification trees for AIC selection of the stock-recruit assumption were the same as those explaining the largest reductions in deviance for the logistic regression models (Figure 3). All branches based on variation in SSB ~~occurred in OMs with larger recruitment variability ($\sigma_R = 1.5$; Figure S17, 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~~ showed better accuracy with larger variability in SSB and all branches based on fishing history showed better accuracy when there was contrast in fishing pressure. Branched based on OM observation error or recruitment variability showed better accuracy when they were lower. For R OMs, a combination of lower recruitment variability, contrast in fishing ~~or variation in natural mortality process error. (Figure S17, 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.~~

pressure, and higher SSB variability produced AIC accuracy over 0.8. 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 S17, C). There was little effect of variability or correlation of selectivity process errors or the level of~~

~~observation error on AIC accuracy .~~

~~Like the~~ S OMs, lower recruitment variability and observation error and higher SSB variability produced AIC accuracy of 0.79. For R+~~Sel OMs, the greatest accuracy for AIC in selecting the Beverton-Holt SRR occurred for~~ M, R+Sel, and R+q ~~OMs where there was contrast in fishing pressure over time which is also where there was the greatest variation in SSB (Figure S17,D). There was also little effect of variability or correlation of catchability process errors or the level of observation error on AIC accuracy~~OM,s accuracy of 0.87 to 0.94 was observed just with higher SSB variability.

Bias

Spawning stock biomass ~~For R OMs ($\sigma_{2+} = 0$), there was no indication of bias (95%confidence intervals included 0) in terminal year SSB~~ Regression models for log-scale errors in SSB that included the various OM and EM factors showed little reduction in deviance ($<$ 4.3%) 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 S6, A). However, errors in terminal SSB estimates were highly variable when~~ factors across all OM process error types (Table 6). The attributes producing the largest reductions were the treatment of median natural mortality ~~was estimated and there was constant fishing pressure and high observation error (Figure S6, A, second row).~~

~~For~~ (for R, 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~~ M, R+~~S OMswith 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 ROMs.~~

~~For R~~Sel, and R+~~M OMswith 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 S6, 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~~q OMs), EM process error type (R+~~M OMs. Like R~~S OMs) and fishing history (all OM types). Including second interactions provided larger further reductions in residual deviance, between 6 and 15%. Including third order interactions also provided large further reductions for R, R+S, and R+~~S OMs , estimating a SRR had little discernible effect on SSB bias~~q OMs (between 5 and 11%).

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~~ generally all regression trees, all branches based on fishing history and level of observation error showed lower bias in SSB with contrast in fishing ~~mortality (Figure S6, 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~~ and lower observation error. ~~However, there was no evidence of SSB bias for correctly specified~~ For branches based on treatment of median natural mortality rate, bias was generally lower when it was known rather than estimated. For some 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 S6, C, third row). We only observed an effect of estimating the~~ and R+q OMs, less bias in SSB was shown when the EM process error configuration was correct.

Stock-recruit parameters

Regression models for transformed relative errors of estimates of both the Beverton-Holt ~~SRR for~~ a and b parameters showed none of the factors explained large percent reductions in deviance (Table 4). The OM fishing history provided the largest reduction for most OM process error types for both parameters, but percent reductions were less than 5.6% except 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 S6, C, top row).~~

~~All EMs fit to data from~~ where the percent reductions were 11.37% and 7.97% for the a and b parameters, respectively and for just the b parameter for 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 S6, D). Many EMs also performed well in R+q OMs with low observation error, but no contrast in fishing pressure.~~ (10%). The EM process error assumption provided similar reductions in deviance for both parameters for R OMs. Including interactions also did not produce important reductions in deviance (increase of .

For regression trees of log-scale error in Beverton-Holt a and b parameters, lower bias was indicated with contrast in OM fishing pressure for all branches in trees for each OM process error type (Figures 4 and 5). For all branches based on recruitment variability in trees 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 S6, 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 S8, 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.~~

OMs, lower bias in both a and b was observed with less recruitment variability. For R OMs with contrast in fishing pressure and greater recruitment variability EMs that assumed the incorrect R+M process errors produced lower bias in both a and b than other process error assumptions. Across all OMs, there was generally less bias and(or) lower variability in 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~~ (Figure S9).

Median natural mortality rate

Fitted regression models for log-scale errors in median natural mortality ~~and a small increase in bias for those OMs that had high observation error (Figure S9, A). For other Rand R~~rate showed largest percent reductions in residual deviance for 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~~ and R_M models (Table 5). The largest reductions for a single attribute was the 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 S9, B) . Relative to the R, and R~~assumption (>20%) and fishing history (>15%) for 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~~ . Fishing history also provided >10% reduction for R+~~Sel OMswith constant fishing pressure (Figure S9, 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 had with the correct process error assumption. There was little effect of estimating natural mortality on relative errors for SRR parameters.~~

~~Like~~ M OMs, but reductions for all factors in R, R+Sel~~OMs, relative errors in SRR parameters for~~ , and R+q OMs were ~~more accurate for most EM process error types when OMs had contrast in fishing pressure and lower observation error (Figure S9, 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~~ relatively low (<6%). Interactions of OM and EM factors also provided substantial further reductions for R+~~q OMs had low observation error and constant fishing pressure (Figure S9, D, fourth row~~S and R+M OMs (between 8 and 15 % for second order interactions).

~~Across all OMs and EMs there was little effect of estimating SRRs on the bias in estimation of~~ Regression trees with branches based on fishing history showed lower bias in median natural mortality ~~(Figure S10). 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 S10, 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~~ rate with contrast in fishing pressure and ~~the EM process error type was correct.~~

branches based on level of observation error showed lower bias with more precise observations (Figure 6). 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 S10, B). However, those R +M OMsthat also had greatest variability in AR1 correlated natural mortality process errors only had unbiased estimation when the~~ OMs, branches based on EM process error ~~type was correct .~~

~~All EM process error types accurately estimated median natural mortality rate for R~~ assumption showed lower bias with EMs assuming the correct R and the incorrect R+~~Sel OMs that had contrast in fishing pressure, low observation error, and low selectivity process error variability (Figure S10, 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~~ S assumption. For R+S and R+~~Sel OMs, all~~ M OMs, branches based on 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 S10, D). Most~~ showed only the correct EM process error ~~types produced biased estimation of median natural mortality when R+q OMs had high observaiton error and constant fishing pressure~~assumption with less bias.

Mohn's ρ

Deviance reduction and regression trees

Fits of glms showed little explanatory power of any of the OM an EM attributes for Mohn's ~~ρ for SSBwas 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 S11, 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,~~ rho for SSB, F, and R. Note that the factors that showed the largest reductions in deviance (which were very small) often do not comport with factors that show differences in median Mohn's ~~ρ was also small in absolute value, but median values were all closer to 0 than the largest valuesin the R and R+S OMs (Figure S11,B-D). For these OMs, there was no noticeable effect of estimation of median natural mortality rate or SRRs on~~ rho results. This is because the glms and regression trees are reducing residual variance and there could be large variability in values, but the medians are less sensitive to extreme values. For example we can see that there is large Mohn's ~~ρ for any EM process error types.~~

~~Mohn's ρ for recruitment was small in absolute value for all~~ rho for R in 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 S13, 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 S13, B-D~~larger variability assumed in recruitment and higher observation error (Figure X), but these factors suggest little reduction in residual deviance (Table X).

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 the 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 (e.g, < 1).

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 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. S17). 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 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. (In review) 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 fitted assessment models have these properties.

On the other hand, estimation of median natural mortality was reliable in many OM scenarios with contrast in fishing pressure, consistent with Hoenig et al. (2025). In some OMs, when EMs estimated the SRR parameters and median natural mortality, 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~~ Li et al. (2024) 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. Contrast in fishing pressure ~~was consistently an important factor across~~ and lower observation error consistently improved all measures of reliability we examined. AIC accurately distinguished models with process errors on recruitment only (R) or on recruitment and apparent survival (R+S). ~~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. ~~AIC generally performed poorly for selecting the SRR, but performance was improved with~~ Accurate AIC selection of the Beverton-Holt stock-recruit relationship generally required a combination of low recruitment variability~~and~~, contrast in fishing pressure ~~. Some~~ and large variation in SSB over time. However, when the Beverton-Holt was correctly assumed, some bias in estimation in at least one of the ~~SRR parameters existed in nearly all OM-EM combinations~~parameters existed regardless of any of the other OM and EM configurations. 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. We thank Jon Deroba, 2 anonymous reviewers, and the associate editor for helpful comments on ~~an earlier version~~ earlier versions of this manuscript that markedly improved its clarity.

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.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. 1984. Classification and regression trees. Chapman; Hall/CRC, New York, NY USA. doi:10.1201/9781315139470.

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. In review. Identification and performance of environmentally-driven stock-recruitment relationships in state space assessment models. Canadian Journal of Fisheries and Aquatic Sciences.

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.

Collier, Z.K., Zhang, H., and Soyoye, O. 2022. Alternative methods for interpreting Monte Carlo experiments. Communications in Statistics - Simulation and Computation: 1–16. doi:10.1080/03610918.2022.2082474.

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.

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. doi:10.1093/icesjms/fsad138.

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. doi:10.1139/cjfas-2012-0112.

Gonzalez, O., O'Rourke, H.P., Wurpts, I.C., and Grimm, K.J. 2018. Analyzing Monte Carlo simulation studies with classification and regression trees. Structural Equation Modeling: A Multidisciplinary Journal **25**(3): 403–413. doi:10.1080/10705511.2017.1369353.

Harwell, M., Kohli, N., and Peralta-Torres, Y. 2018. A survey of reporting practices of computer simulation studies in statistical research. The American Statistician **72**(4): 321–327. doi:10.1080/00031305.2017.1342692.

Hoenig, J.M., Hearn, W.S., Leigh, G.M., and Latour, R.J. 2025. Principles for estimating natural mortality rate. Fisheries Research **281**: 107195. doi:10.1016/j.fishres.2024.107195.

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. doi:10.1016/j.fishres.2021.106176.

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.

Johnson, K.F., Councill, E., Thorson, J.T., Brooks, E., Methot, R.D., and Punt, A.E. 2016. Can autocorrelated recruitment be estimated using integrated assessment models and how does it affect population forecasts? Fisheries Research **183**: 222–232. doi:10.1016/j.fishres.2016.06.004.

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. doi:10.18637/jss.v070.i05.

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. doi:10.1139/cjfas-2015-0232.

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 account for movement dynamics and improve 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 review. Guidance on bias-correction of log-normal random effects and observations in state-space assessment models. Canadian Journal of Fisheries and Aquatic Sciences.

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. doi:10.1016/j.fishres.2024.107023.

MacKinnon, D.P., Warsi, G., and Dwyer, J.H. 1995. A simulation study of mediated effect measures. Multivariate Behavioral Research **30**(1): 41–62. doi:10.1207/s15327906mbr3001_3.

Magnusson, A., and Hilborn, R. 2007. What makes fisheries data informative? Fish and Fisheries **8**(4): 337–358. doi: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_

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.

Pontavice, H. du, 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.

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. doi:10.1016/j.fishres.2023.106642.

Punt, A.E., Hurtado-Ferro, F., and Whitten, A.R. 2014. Model selection for selectivity in fisheries stock assessments. Fisheries Research **158**: 124–134. doi:10.1016/j.fishres.2013.06.003.

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. doi:10.1016/j.fishres.2016.06.018.

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: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. doi:10.1016/j.fishres.2021.105873.

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. doi:10.1111/1365-2664.13515.

Wang, S., Cadigan, N.G., and Benoît, H.P. 2017. Inference about regression parameters using highly stratified survey count data with over-dispersion and repeated measurements. Journal of Applied Statistics **44**(6): 1013–1030. doi:10.1080/02664763.2016.1191622.

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:10.1016/j.fishres.2018.09.018.

Xu, H., Thorson, J.T., Methot, R.D., and Taylor, I.G. 2019. A new semi-parametric method for autocorrelated age- and time-varying selectivity in age-structured assessment models. Canadian Journal of Fisheries and Aquatic Sciences **76**(2): 268–285. doi:10.1139/cjfas-2017-0446.

Yee, T.W. 2008. The VGAM package. R News **8**(2): 28–39. Available from https://journal.r-project.org/articles/RN-2008-014/.

Yee, T.W. 2015. Vector generalized linear and additive models: With an implementation in R. Springer, New York, NY USA. doi:10.1007/978-1-4939-2818-7.

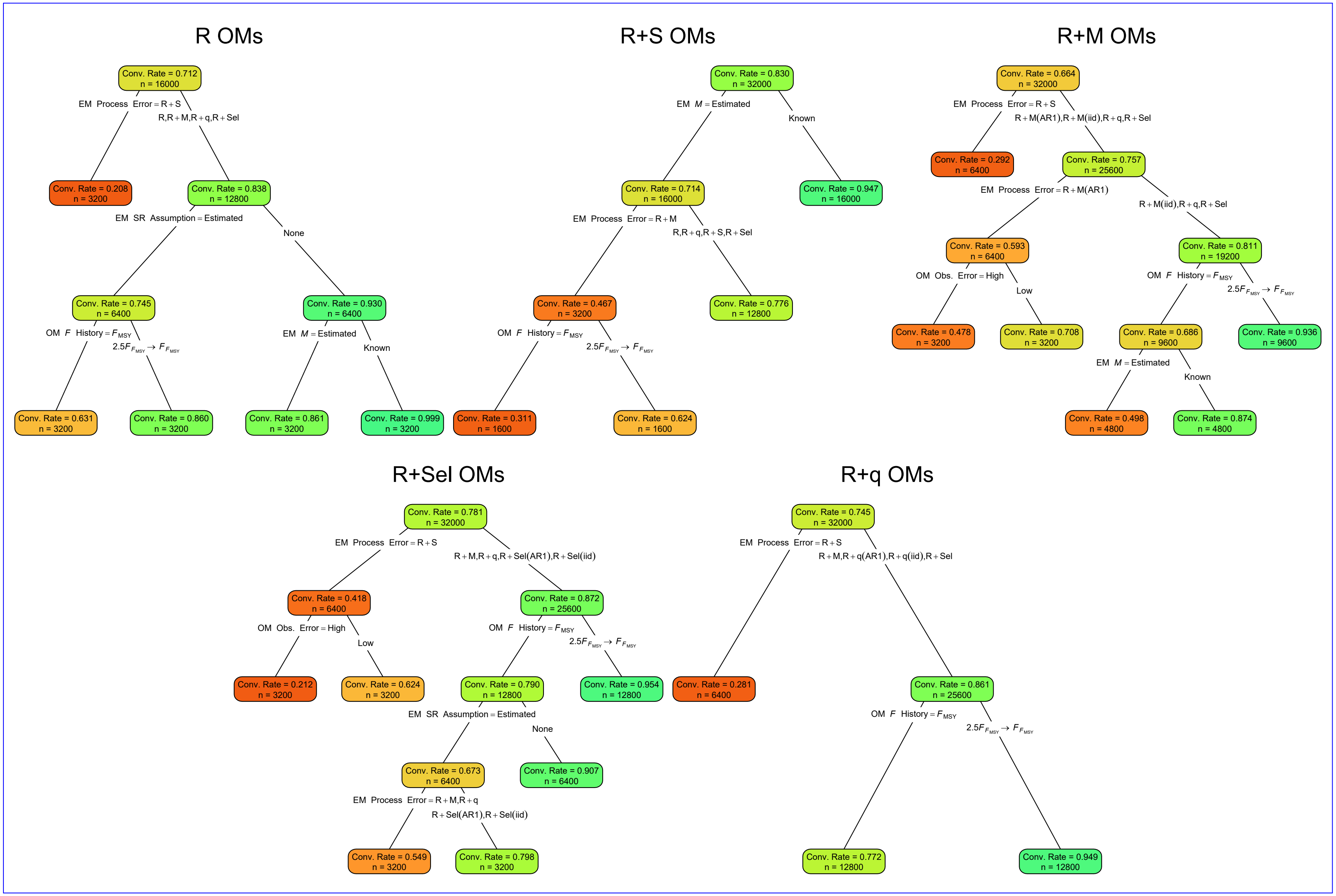

Fig. 1. ~~Estimated probability of fits~~ Classification trees indicating primary factors determining convergence as defined by providing hessian-based standard errors for ~~EMs assuming alternative process error (colored points and lines)~~R, ~~and median natural mortality (estimated or known) and Beverton-Holt stock-recruit relationships (estimated or not; along x-axis) when fitted to operating models that have~~ R~~and R~~+S~~(A)~~, R+~~Sel (B)~~M, R+~~M (C), or~~ Sel and R+q ~~(D) process error structures~~OMs. ~~Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.~~Lower or higher convergence rates are indicated by more red or green polygons, respectively

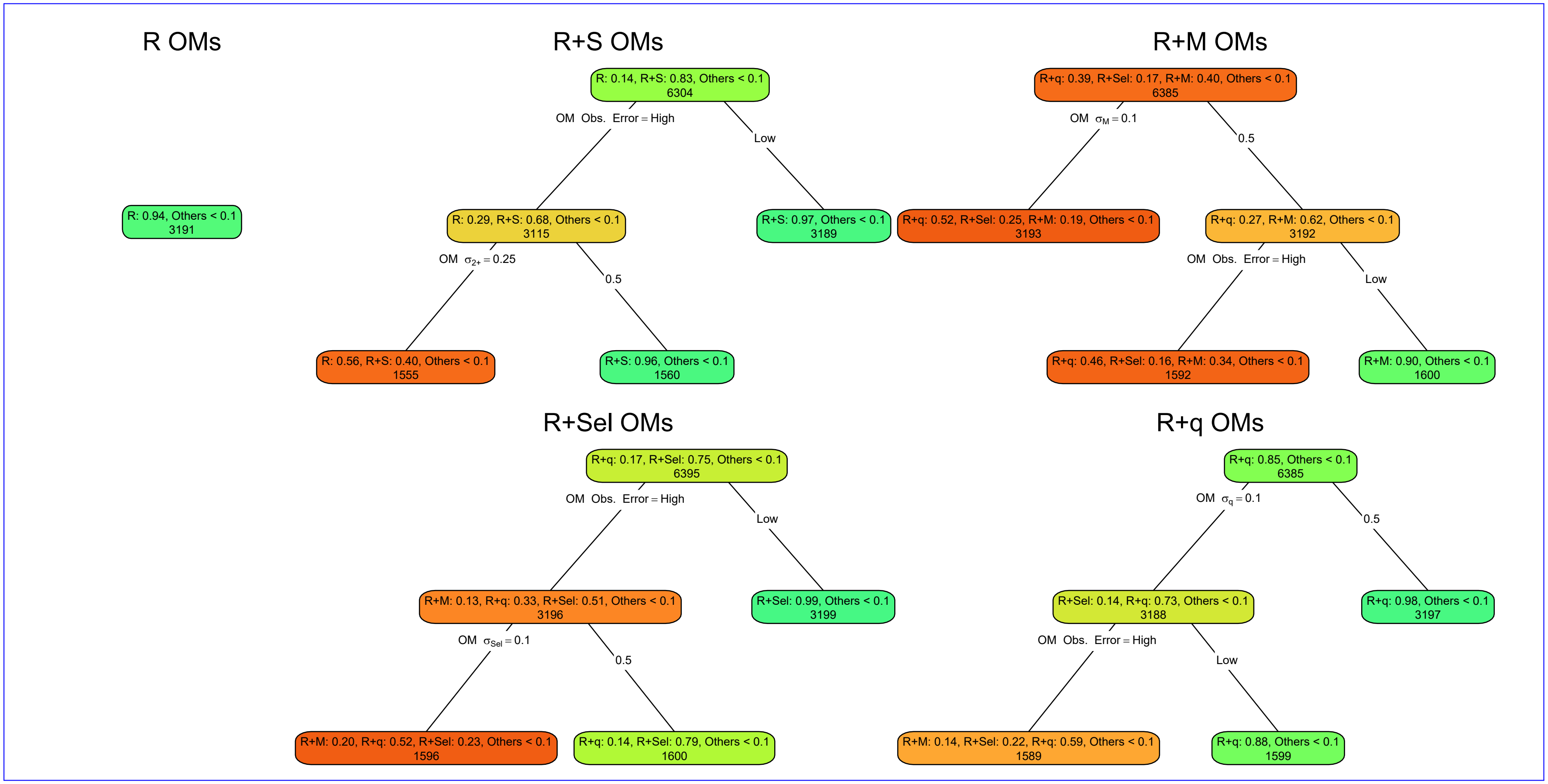

Fig. 2. Classification trees indicating primary factors determining which EM process error assumption provides the lowest AIC for R+S, R+M, R+Sel and R+q OMs. Each node shows the proportion of EM process error models with lowest AIC (top) and number of observations (bottom) for the corresponding subset. Lower or higher accuracy of the process error assumption are indicated by more red or green polygons, respectively.

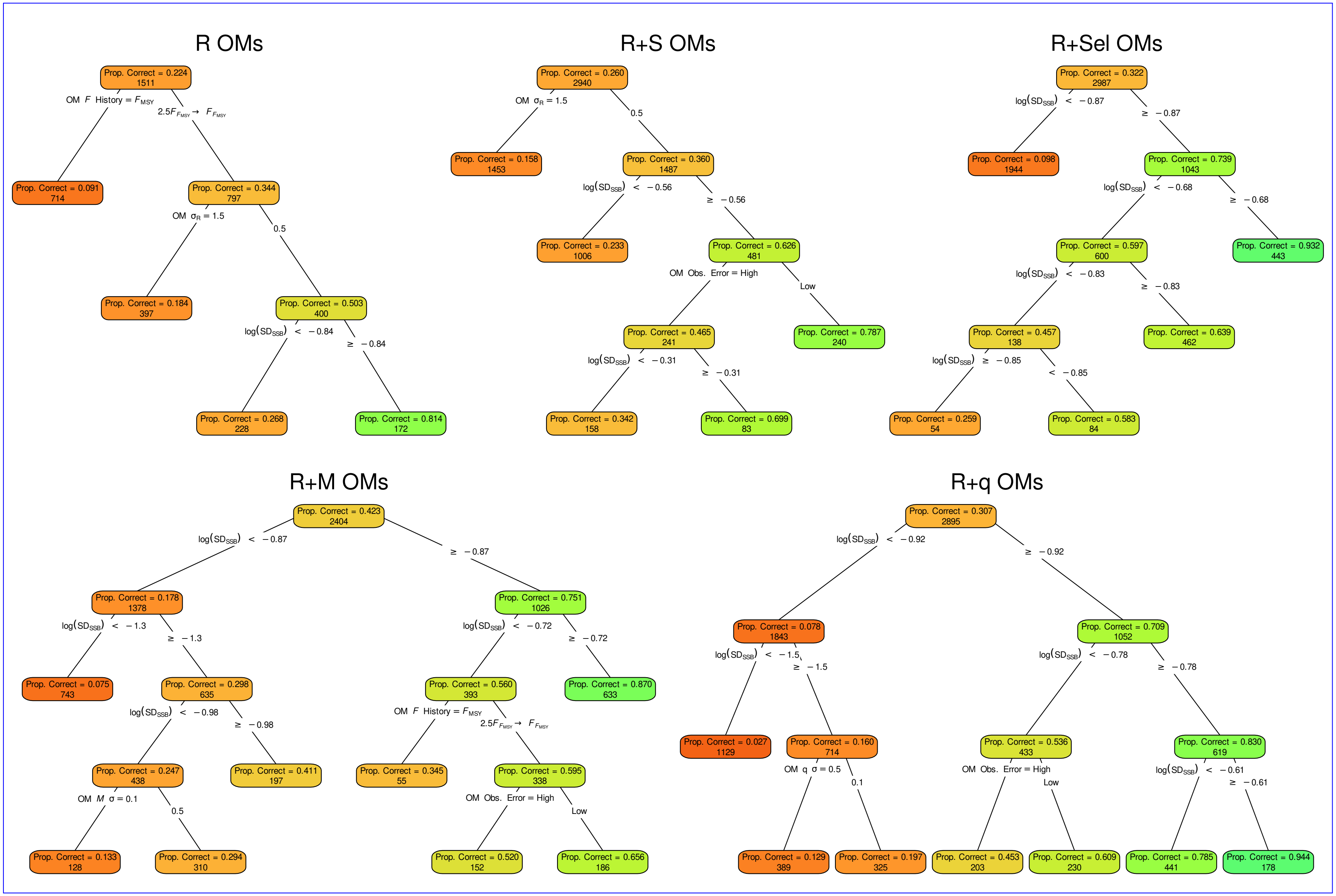

Fig. 3. ~~Estimated probability of lowest AIC for EMs assuming alternative process error structures~~ Classification trees indicating primary factors determining which EM stock recruitment assumption (~~colored bars) conditional on alternative assumptions for median natural mortality (estimated~~ none or ~~known) and~~ Beverton-Holt~~stock-recruit relationships (estimated or not; along x-axis~~) ~~when fitted to operating models that have~~ provides the lowest AIC for R~~and~~, R+S~~(A)~~, R+~~Sel (B)~~M, R+~~M (C), or~~ Sel and R+q OMs. Each node shows the proportion of EMs that assume the stock-recruit relationship with lowest AIC (~~D~~top) ~~process error structures~~and number of observations (bottom) for the corresponding subset. ~~Striped bars indicate results where~~

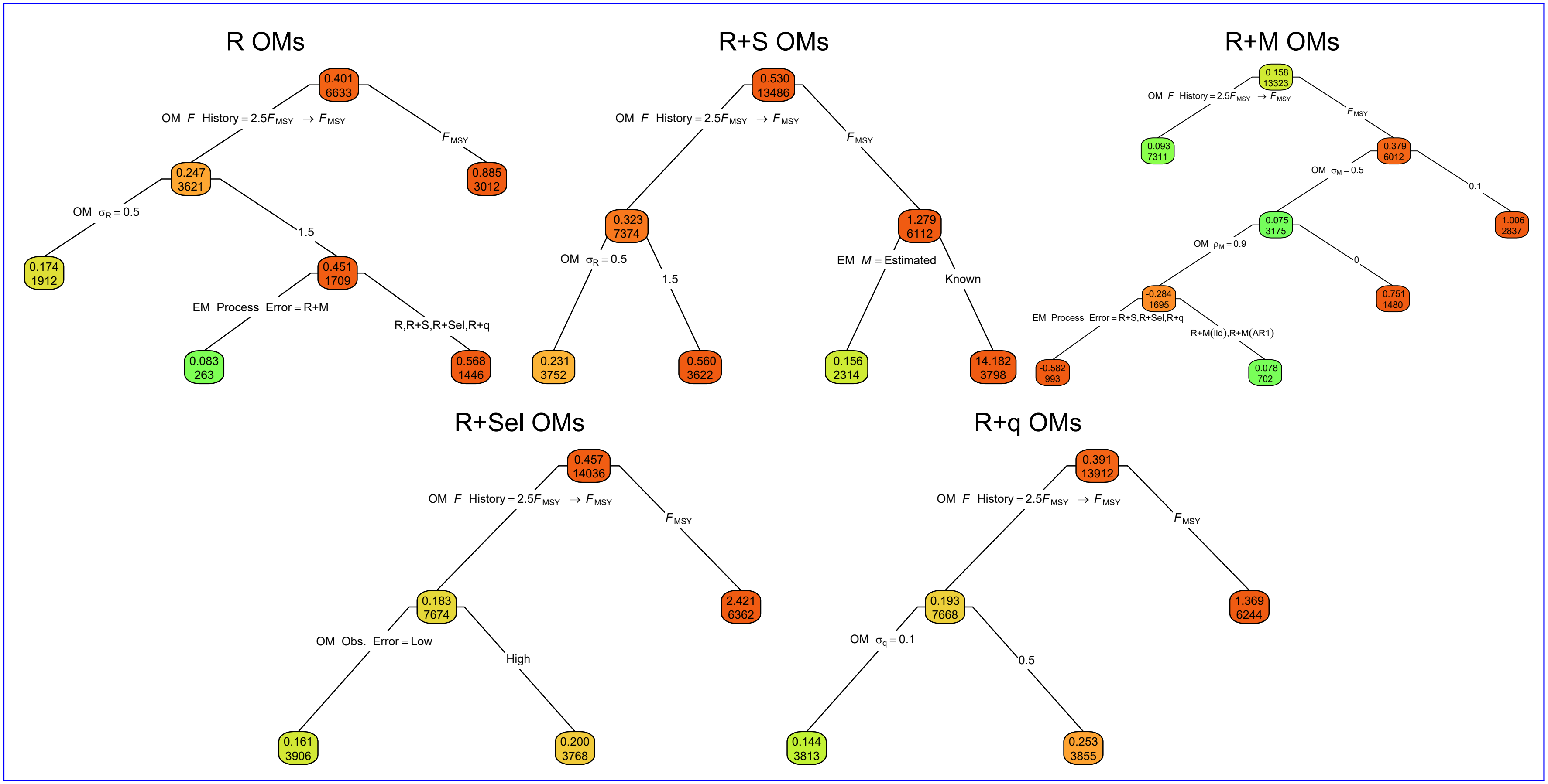

Fig. 4. Regression trees indicating primary factors determining reductions in sums of squares of errors in estimation measured by Eq. 3 for the Beverton-Holt stock-recruit parameter a for R+S, R+M, R+Sel and R+q OMs. Each node shows the median error (top) and number of observations (bottom) for the corresponding subset. Lower or higher median absolute errors of the process error assumption are indicated by more green or red polygons, respectively.

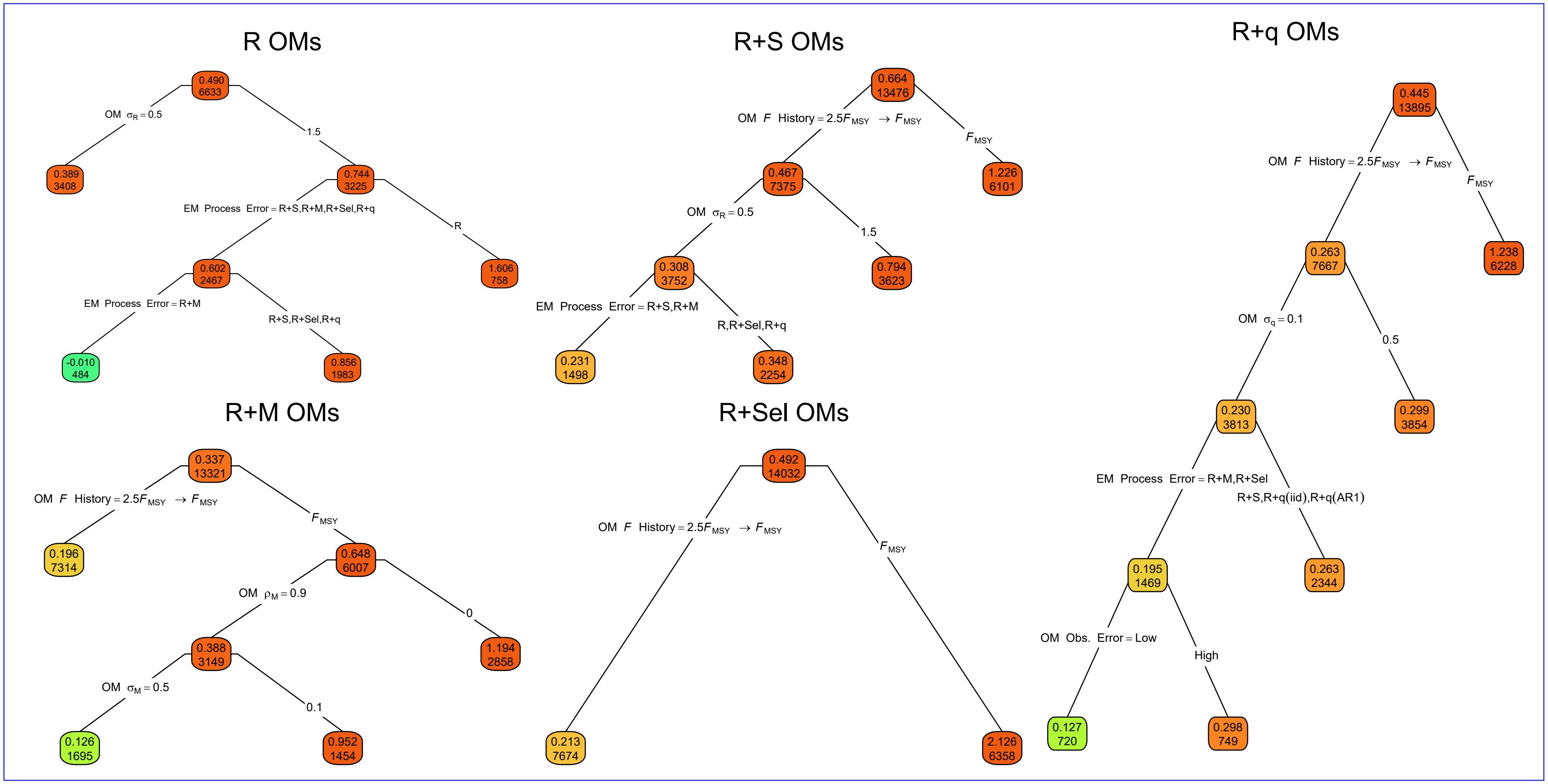

Fig. 5. ~~Estimated probability~~ Regression trees indicating primary factors determining reductions in sums of ~~lowest AIC from logistic regression on the log-standard deviation~~ squares of ~~the true log(SSB)~~ errors in ~~each simulation~~ estimation measured by Eq. 3 for ~~estimating model with~~ the 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.~~ parameter b for R~~and R~~+S~~(A)~~, R+~~Sel (B)~~M, R+~~M (C), or~~ Sel and R+q OMs. Each node shows the median error (~~D~~top) ~~,~~ 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~~ number of ~~results indicates the range of log-standard deviation of log~~observations (~~SSB~~bottom) for ~~simulations~~ the corresponding subset. Lower or higher median absolute errors of the ~~particular OM~~process error assumption are indicated by more green or red polygons, respectively.

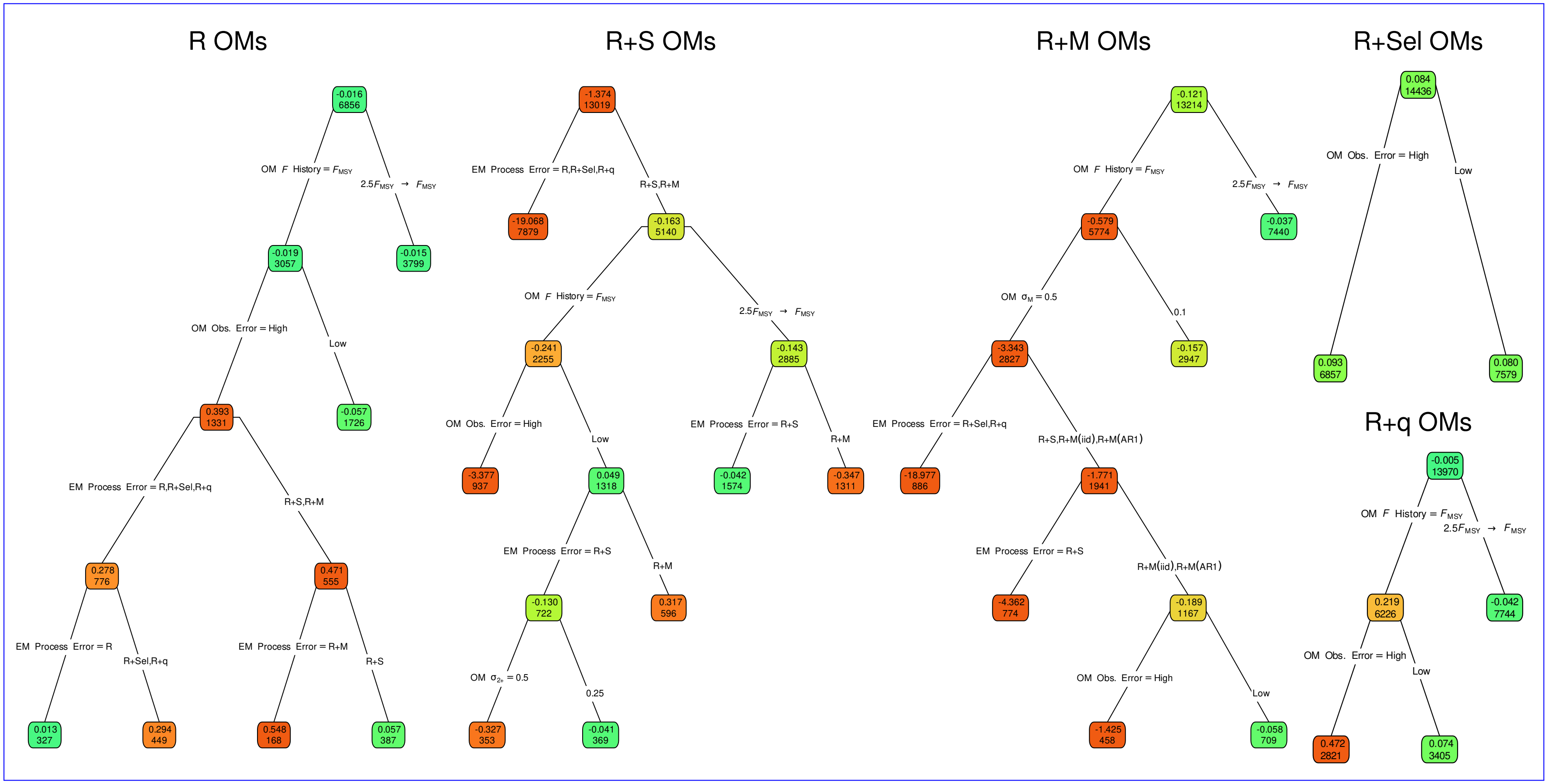

Fig. 6. ~~Median relative error~~ Regression trees indicating primary factors determining reductions in sums of ~~terminal year SSB~~ squares of errors in estimation measured by Eq. 3 for ~~estimating models fitted to data sets simulated with alternative process error structures:~~ the median natural mortality rate for R~~and R~~+S~~(A)~~, R+~~Sel (B)~~M, R+~~M (C), or~~ Sel and R+q OMs. Each node shows the median error (~~D~~top) and number of observations (bottom) for the corresponding subset. ~~Circled values indicate results where~~ Lower or higher median absolute errors of the ~~EM~~ process error ~~structure matches that of the operating model and vertical lines represent 95% confidence intervals~~assumption are indicated by more green or red polygons, respectively.

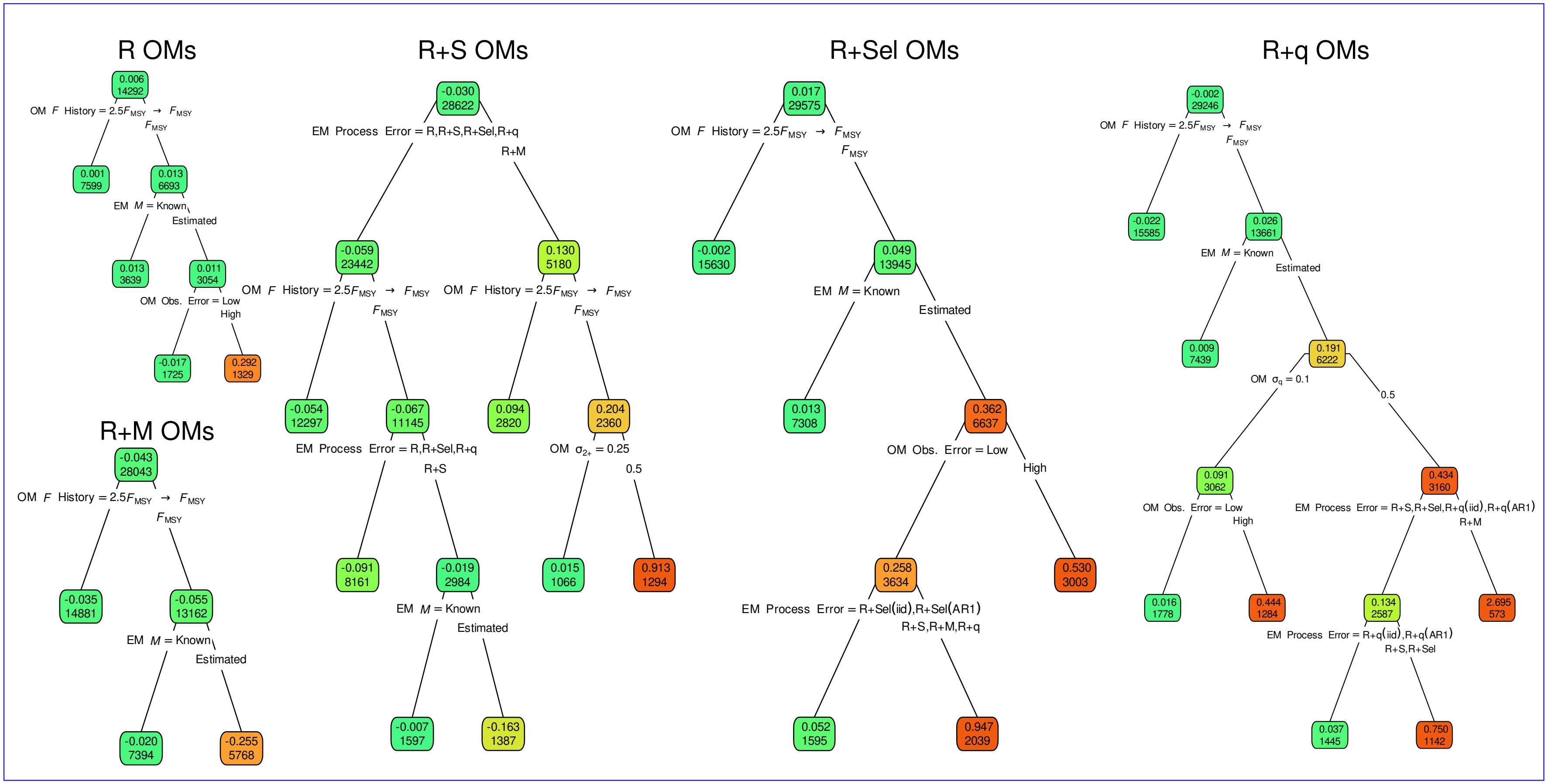

Fig. 7. ~~Median Mohn's rho~~ Regression trees indicating primary factors determining reductions in sums of squares of errors in estimation measured by Eq. 3 for ~~SSB~~ terminal year spawning stock biomass for ~~estimating models fitted to data sets simulated with alternative process error structures:~~ R~~and R~~+S~~(A)~~, R+~~Sel (B)~~M, R+~~M (C), or~~ Sel and R+q ~~(D)~~OMs. ~~Circled values indicate results where~~ Each node shows the ~~EM process~~ median error ~~structure matches that~~ (top) and number of observations (bottom) for the ~~operating model and vertical lines represent 95% confidence intervals~~corresponding subset. Median errors closer to or further from zero are indicated by more green or red polygons, respectively.

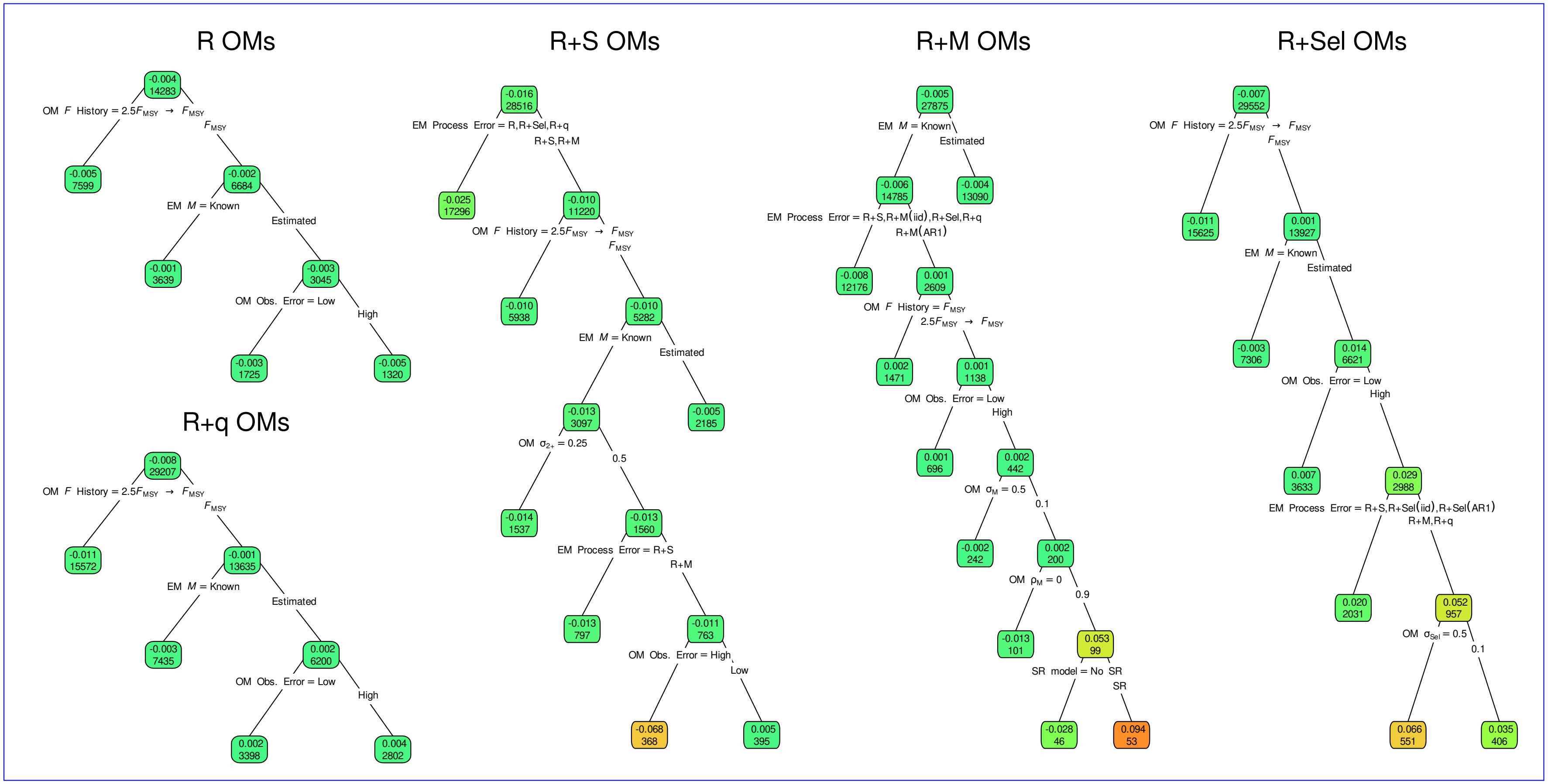

Fig. 8. Regression trees indicating primary factors determining reductions in sums of squares of errors in transformed Mohn's ρ (Eq. 3) for spawning stock biomass for R+S, R+M, R+Sel and R+q OMs. Each node shows the median Mohn's ρ (top) and number of observations (bottom) for the corresponding subset. Median Mohn's ρ closer to or further from zero are indicated by more green or red polygons, respectively.

Table 1. For each OM process error type (columns), percent reduction in deviance for logistic regression models fit to indicators of convergence (providing hessian-based standard errors) with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM Process Error	27.95	4.58	14.68	17.24	24.66
EM M Assumption	1.07	11.43	2.45	0.56	1.46
EM SR Assumption	2.88	3.30	1.24	2.47	1.59
OM Obs. Error	0.75	4.64	2.06	4.54	1.60
OM F History	2.32	3.37	1.63	3.30	2.59
OM σ_R	0.10	0.02	–	–	–
OM σ_{2+}	–	0.40	–	–	–
OM σ_M	–	–	0.22	–	–
OM ρ_R	–	–	0.17	–	–
OM σ_{Sel}	–	–	–	1.81	–
OM ρ_{Sel}	–	–	–	0.02	–
OM σ_q	–	–	–	–	0.34
OM ρ_q	–	–	–	–	<0.01
All factors	39.54	31.46	24.85	34.83	36.31
+ All Two Way	45.03	39.89	35.20	42.81	43.70
+ All Three Way	47.02	44.57	37.88	45.51	46.87

Table 2. For each OM process error type (columns), percent reduction in deviance for multinomial logistic regression models fit to indicators of EM process error assumption with lowest AIC with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	5.52	1.05	0.52	0.61	1.32
EM SR Assumption	5.60	0.75	1.13	0.93	1.95
OM Obs. Error	2.96	22.46	3.42	25.67	5.03
OM F History	5.77	0.62	0.94	0.91	2.05
OM σ_R	0.10	0.66	–	–	–
OM σ_{2+}	–	16.86	–	–	–
OM σ_M	–	–	9.06	–	–
OM ρ_R	–	–	0.38	–	–
OM σ_{Sel}	–	–	–	7.59	–
OM ρ_{Sel}	–	–	–	0.60	–
OM σ_q	–	–	–	–	13.50
OM ρ_q	–	–	–	–	0.75
All factors	20.98	46.12	16.58	40.83	25.99
+ All Two Way	22.02	48.94	21.63	44.08	30.17
+ All Three Way	22.05	49.98	22.36	44.54	31.38

Table 3. For each OM process error type (columns), percent reduction in deviance for logistic regression models fit to indicators of EM stock-recruit assumption (none or Beverton-Holt) with lowest AIC with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	0.04	0.21	0.18	0.02	0.01
OM Obs. Error	<0.01	0.65	0.14	0.04	0.02
OM F History	9.17	3.79	13.08	26.56	24.60
OM σ_R	3.54	4.74	–	–	–
OM σ_{2+}	–	0.14	–	–	–
OM σ_M	–	–	1.14	–	–
OM ρ_R	–	–	0.05	–	–
OM σ_{Sel}	–	–	–	0.02	–
OM ρ_{Sel}	–	–	–	0.17	–
OM σ_q	–	–	–	–	0.36
OM ρ_q	–	–	–	–	0.02
log (SD$_{\mathrm{SSB}}$)	4.11	1.59	33.39	41.36	39.23
All factors	31.52	18.99	34.23	43.77	42.31
+ All Two Way	34.79	22.24	35.99	45.84	44.04
+ All Three Way	35.41	23.09	37.57	46.39	44.63

Table 4. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to errors in estimation measured by Eq. 3 for the Beverton-Holt stock recruit relation ship parameters with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	**Beverton-Holt** a					**Beverton-Holt** b				
	R	R+S	R+M	R+Sel	R+q	R	R+S	R+M	R+Sel	R+q
EM M Assumption	0.02	1.05	0.02	0.11	0.02	0.05	1.06	0.03	0.01	0.40
EM Process Error	2.74	0.18	0.20	1.25	1.90	2.29	1.21	0.12	1.40	3.06
OM Obs. Error	0.16	$<$0.01	0.01	0.04	$<$0.01	$<$0.01	0.01	0.05	0.01	0.01
OM F History	3.15	3.34	5.60	11.37	10.00	1.16	1.17	2.01	7.97	3.87
OM σ_R	2.31	0.74	–	–	–	1.67	0.52	–	–	–
OM σ_{2+}	–	0.29	–	–	–	–	0.01	–	–	–
OM σ_M	–	–	0.30	–	–	–	–	0.13	–	–
OM ρ_R	–	–	0.51	–	–	–	–	0.22	–	–
OM σ_{Sel}	–	–	–	0.13	–	–	–	–	0.05	–
OM ρ_{Sel}	–	–	–	0.07	–	–	–	–	0.04	–
OM σ_q	–	–	–	–	0.04	–	–	–	–	0.10
OM ρ_q	–	–	–	–	$<$0.01	–	–	–	–	$<$0.01
All factors	8.07	5.15	6.73	12.64	11.79	4.91	3.75	2.55	9.12	7.22
+ All Two Way	9.96	7.37	9.76	13.59	13.65	7.55	7.15	4.32	10.08	12.16
+ All Three Way	11.22	8.15	11.13	14.48	14.87	9.78	9.02	5.26	11.08	14.73

Table 5. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to errors in estimation measured by Eq. 3 for the median natural mortality rate parameter with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM SR assumption	0.21	0.38	0.11	0.26	0.43
EM Process Error	1.98	20.36	3.16	0.94	1.31
OM Obs. Error	4.74	0.79	0.40	2.23	1.88
OM F History	5.07	15.11	10.65	0.24	2.38
OM σ_R	<0.01	0.01	–	–	–
OM σ_{2+}	–	5.04	–	–	–
OM σ_M	–	–	5.32	–	–
OM ρ_R	–	–	0.85	–	–
OM σ_{Sel}	–	–	–	1.30	–
OM ρ_{Sel}	–	–	–	0.37	–
OM σ_q	–	–	–	–	0.46
OM ρ_q	–	–	–	–	0.06
All factors	12.64	40.10	21.29	5.54	6.52
+ All Two Way	21.17	48.12	36.19	9.87	11.71
+ All Three Way	23.03	50.38	42.82	11.58	14.64

Table 6. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to errors in estimation measured by Eq. 3 for the terminal year spawning biomass with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	2.28	1.15	1.04	2.92	3.26
EM SR assumption	0.10	0.06	0.08	0.06	0.08
EM Process Error	0.43	4.28	0.40	0.11	1.05
OM Obs. Error	1.63	0.07	0.78	0.32	<0.01
OM F History	2.62	3.15	1.28	3.22	4.72
OM σ_R	0.03	0.01	–	–	–
OM σ_{2+}	–	0.93	–	–	–
OM σ_M	–	–	0.18	–	–
OM ρ_R	–	–	0.01	–	–
OM σ_{Sel}	–	–	–	0.16	–
OM ρ_{Sel}	–	–	–	0.04	–
OM σ_q	–	–	–	–	1.02
OM ρ_q	–	–	–	–	0.06
All factors	7.59	9.86	3.93	7.04	10.64
+ All Two Way	17.99	25.56	10.06	13.44	22.43
+ All Three Way	23.39	36.74	13.76	16.55	31.11

Table 7. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to transformed Mohn's ρ values for each simulation (Eq. 3) for spawning biomass with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	0.79	0.18	0.15	0.95	1.24
EM SR assumption	<0.01	0.01	<0.01	<0.01	<0.01
EM Process Error	<0.01	0.22	0.14	0.08	0.04
OM Obs. Error	0.12	0.03	0.05	0.18	0.21
OM F History	0.84	0.14	0.07	1.08	1.56
OM σ_R	0.01	0.01	–	–	–
OM σ_{2+}	–	0.02	–	–	–
OM σ_M	–	–	0.01	–	–
OM ρ_R	–	–	<0.01	–	–
OM σ_{Sel}	–	–	–	0.01	–
OM ρ_{Sel}	–	–	–	0.02	–
OM σ_q	–	–	–	–	0.01
OM ρ_q	–	–	–	–	0.01
All factors	1.89	0.63	0.43	2.43	3.29
+ All Two Way	3.63	1.10	0.91	4.75	6.22
+ All Three Way	4.27	1.65	1.50	5.73	7.53

Supplementary Materials

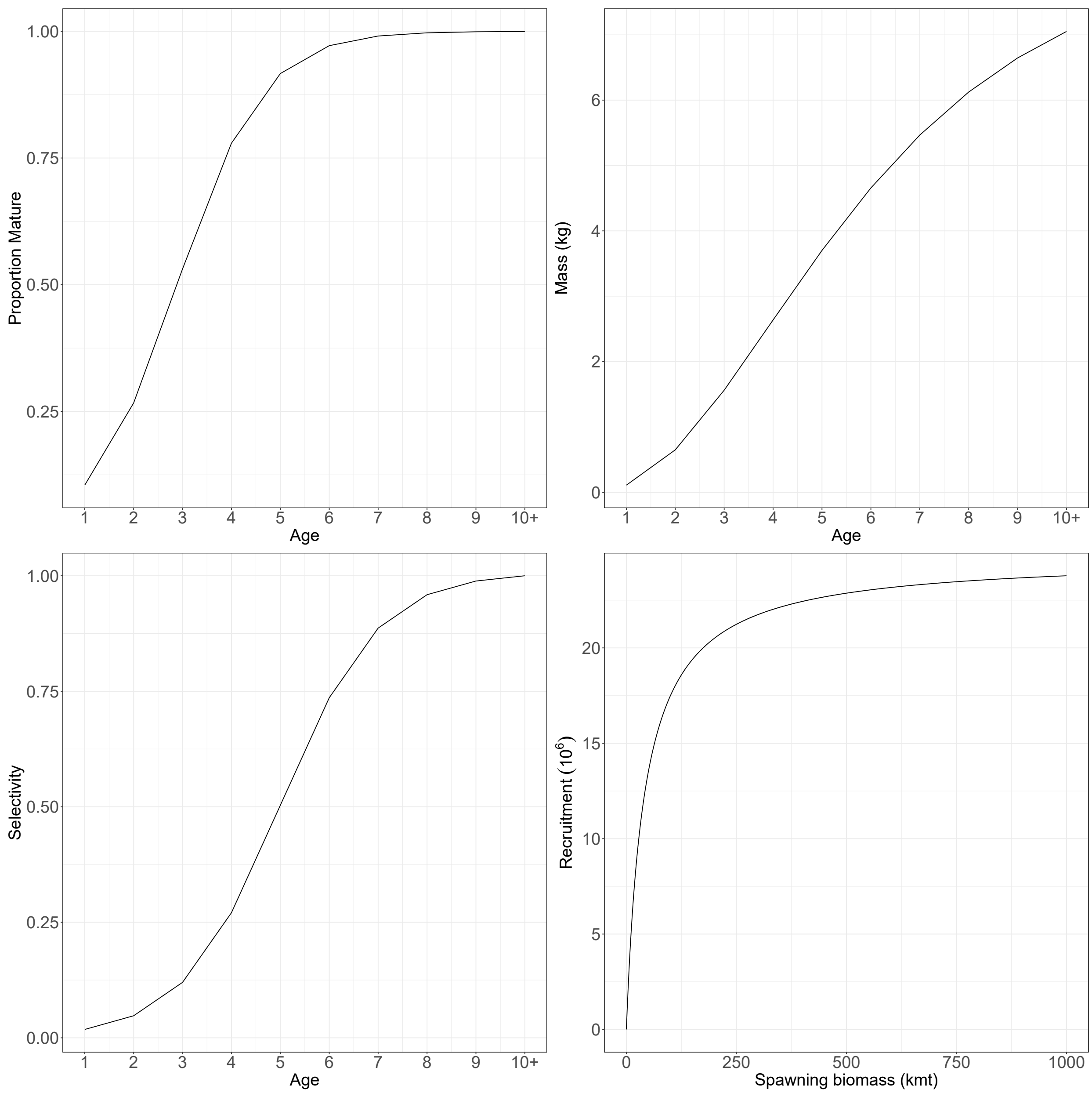

Fig. S1. The proportion mature at age, weight at age, fleet and index selectivity at age, and Beverton-Holt stock-recruit relationship assumed for the population in all operating models. For operating models with random effects on fleet selectivity, this represents the selectivity at the mean of the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

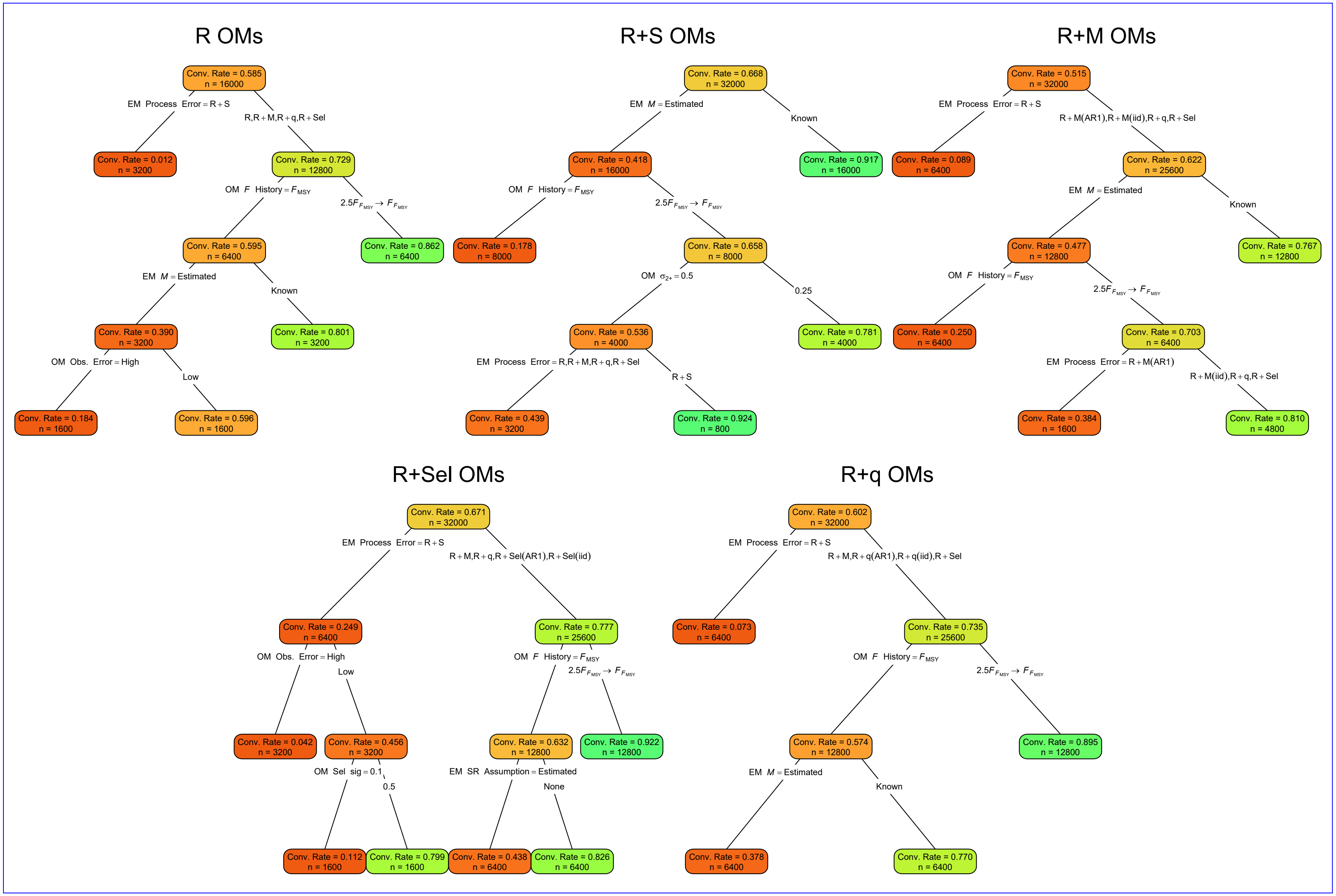

Fig. S2. Classification trees indicating primary factors determining convergence as defined by a maximum absolute gradient $< 10^{-6}$ for R, R+S, R+M, R+Sel and R+q OMs. Lower or higher convergence rates are indicated by more red or green polygons, respectively

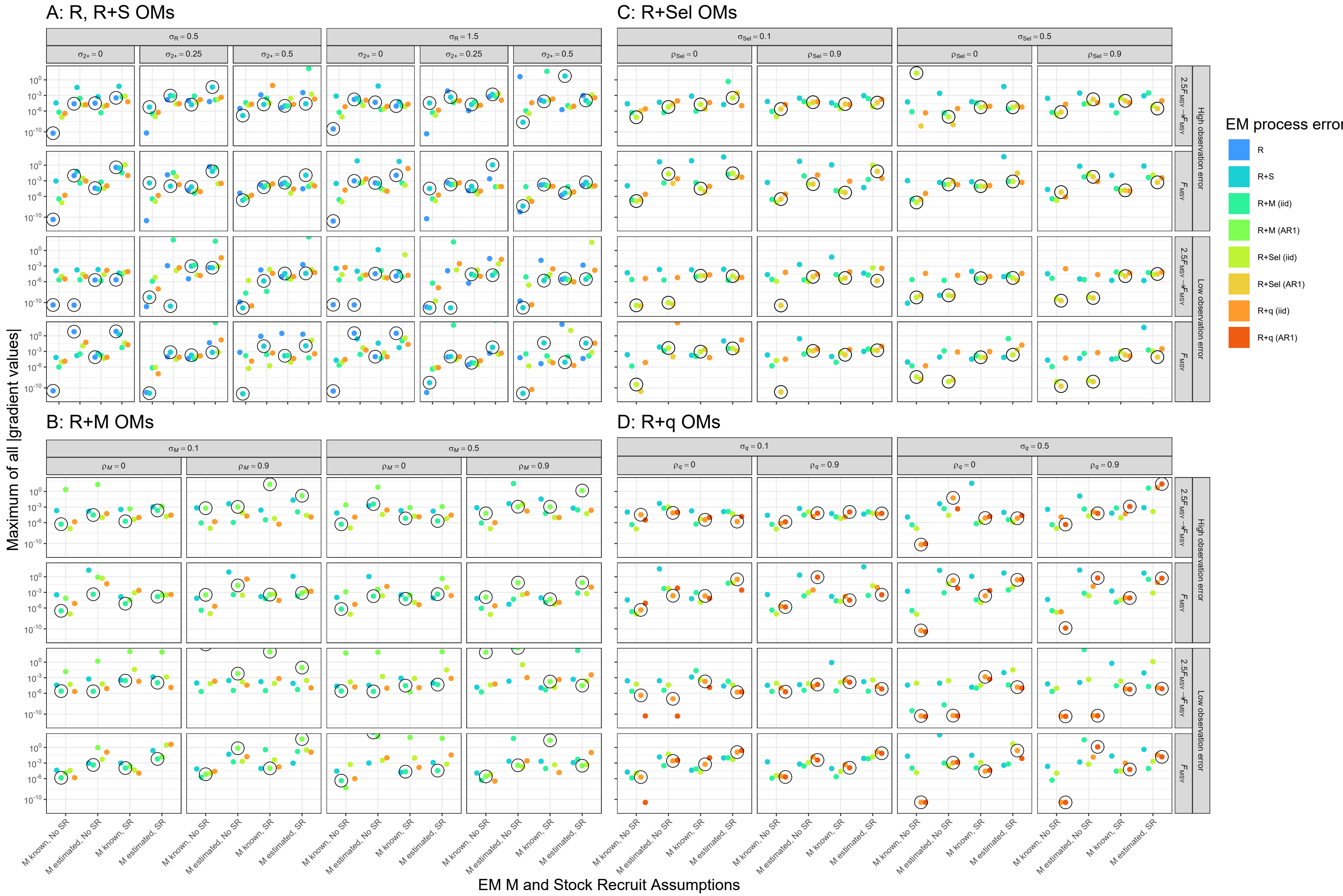

Fig. S3. The maximum of the absolute values of all gradient values for all fits that provided hessian-based standard errors across all simulated data sets of a given OM configuration (A: R and R+S, B: R+M, C: R+Sel, or D: R+q). Results are conditional on EM fits with alternative process error type (colored points and lines), median natural mortality (estimated or known) and recruitment assumptions (Beverton-Holt stock-recruit 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.

Table S1. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to errors in estimation measured by Eq. 3 for the terminal year fully-selected fishing mortality with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	2.26	1.33	1.26	2.93	3.26
EM SR assumption	0.11	0.07	0.08	0.07	0.09
EM Process Error	0.46	4.18	0.38	0.13	1.02
OM Obs. Error	1.61	0.06	0.86	0.41	<0.01
OM F History	2.49	3.23	1.42	3.22	4.55
OM σ_R	0.02	0.02	–	–	–
OM σ_{2+}	–	0.87	–	–	–
OM σ_M	–	–	0.16	–	–
OM ρ_R	–	–	0.01	–	–
OM σ_{Sel}	–	–	–	0.24	–
OM ρ_{Sel}	–	–	–	0.05	–
OM σ_q	–	–	–	–	1.03
OM ρ_q	–	–	–	–	0.05
All factors	7.42	9.96	4.37	7.26	10.43
+ All Two Way	17.63	25.76	10.94	13.88	22.07
+ All Three Way	22.97	37.03	14.74	17.32	30.74

Table S2. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to errors in estimation measured by Eq. 3 for the terminal year recruitment with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	1.96	0.40	0.69	3.52	3.03
EM SR assumption	0.06	0.02	0.05	0.02	0.05
EM Process Error	0.39	4.74	0.41	0.12	1.16
OM Obs. Error	1.47	0.08	0.64	0.18	<0.01
OM F History	2.54	2.66	1.11	4.18	5.06
OM σ_R	0.03	0.01	–	–	–
OM σ_{2+}	–	1.05	–	–	–
OM σ_M	–	–	0.36	–	–
OM ρ_R	–	–	0.02	–	–
OM σ_{Sel}	–	–	–	0.23	–
OM ρ_{Sel}	–	–	–	0.06	–
OM σ_q	–	–	–	–	1.09
OM ρ_q	–	–	–	–	0.06
All factors	6.90	9.01	3.43	8.58	10.90
+ All Two Way	16.48	24.64	9.73	15.76	22.75
+ All Three Way	21.46	35.60	13.56	19.07	31.15

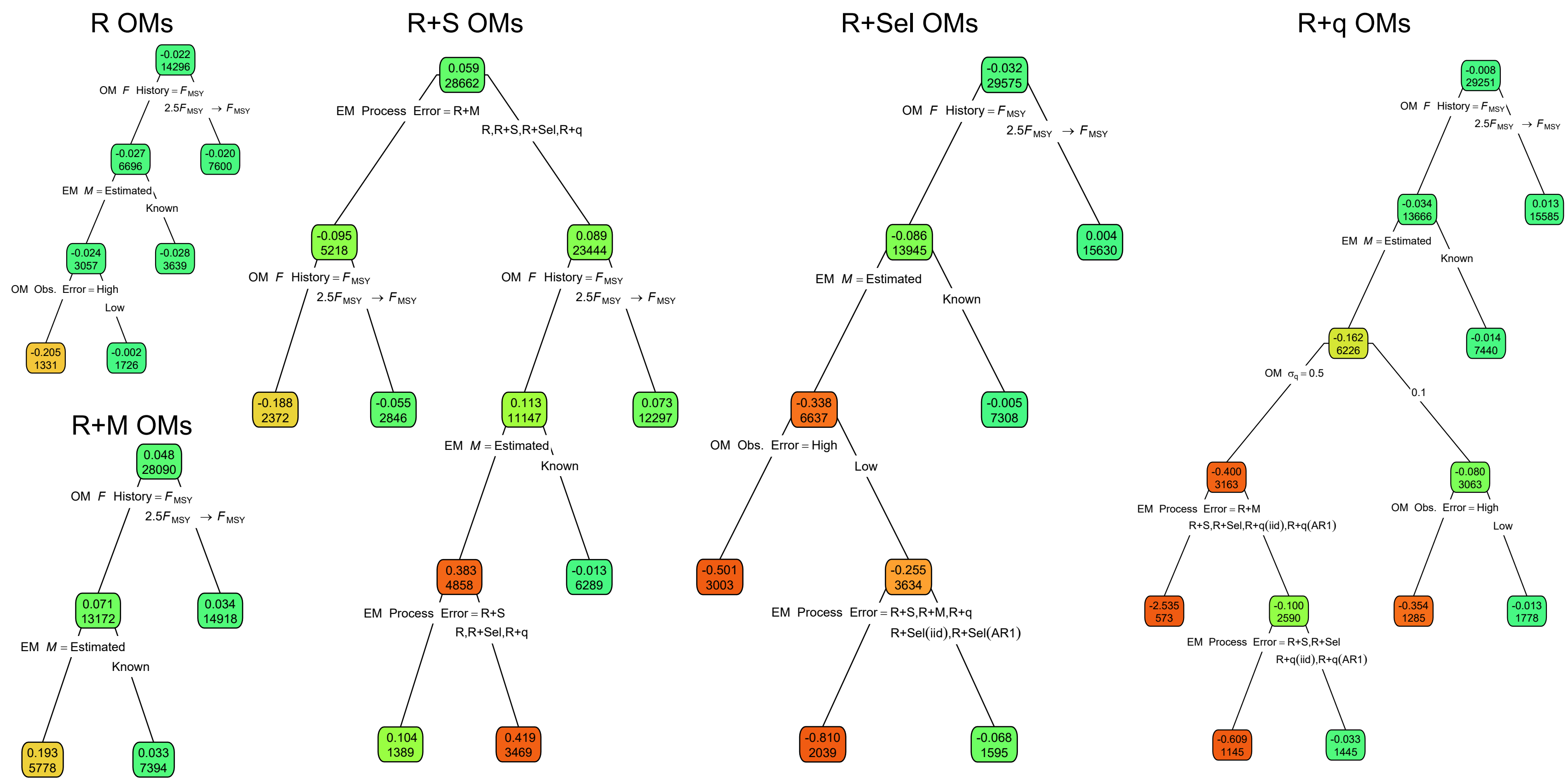

Fig. S4. Regression trees indicating primary factors determining reductions in sums of squares of errors in estimation measured by Eq. 3 for terminal year fully-selected fishing mortality for R+S, R+M, R+Sel and R+q OMs. Each node shows the median error (top) and number of observations (bottom) for the corresponding subset. Median errors closer to or further from zero are indicated by more green or red polygons, respectively.

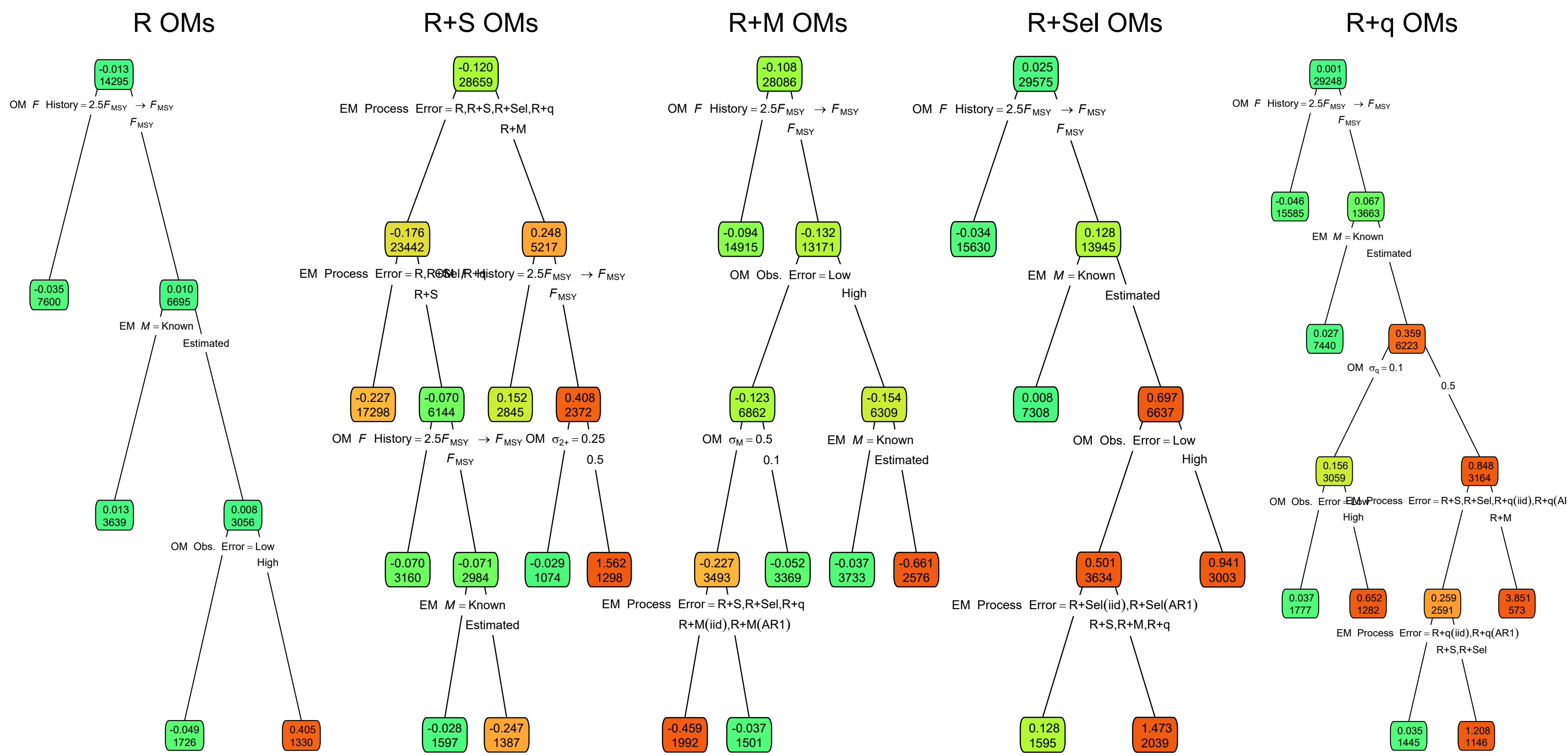

Fig. S5. Regression trees indicating primary factors determining reductions in sums of squares of errors in estimation measured by Eq. 3 for terminal year recruitment for R+S, R+M, R+Sel and R+q OMs. Each node shows the median error (top) and number of observations (bottom) for the corresponding subset. Median errors closer to or further from zero are indicated by more green or red polygons, respectively.

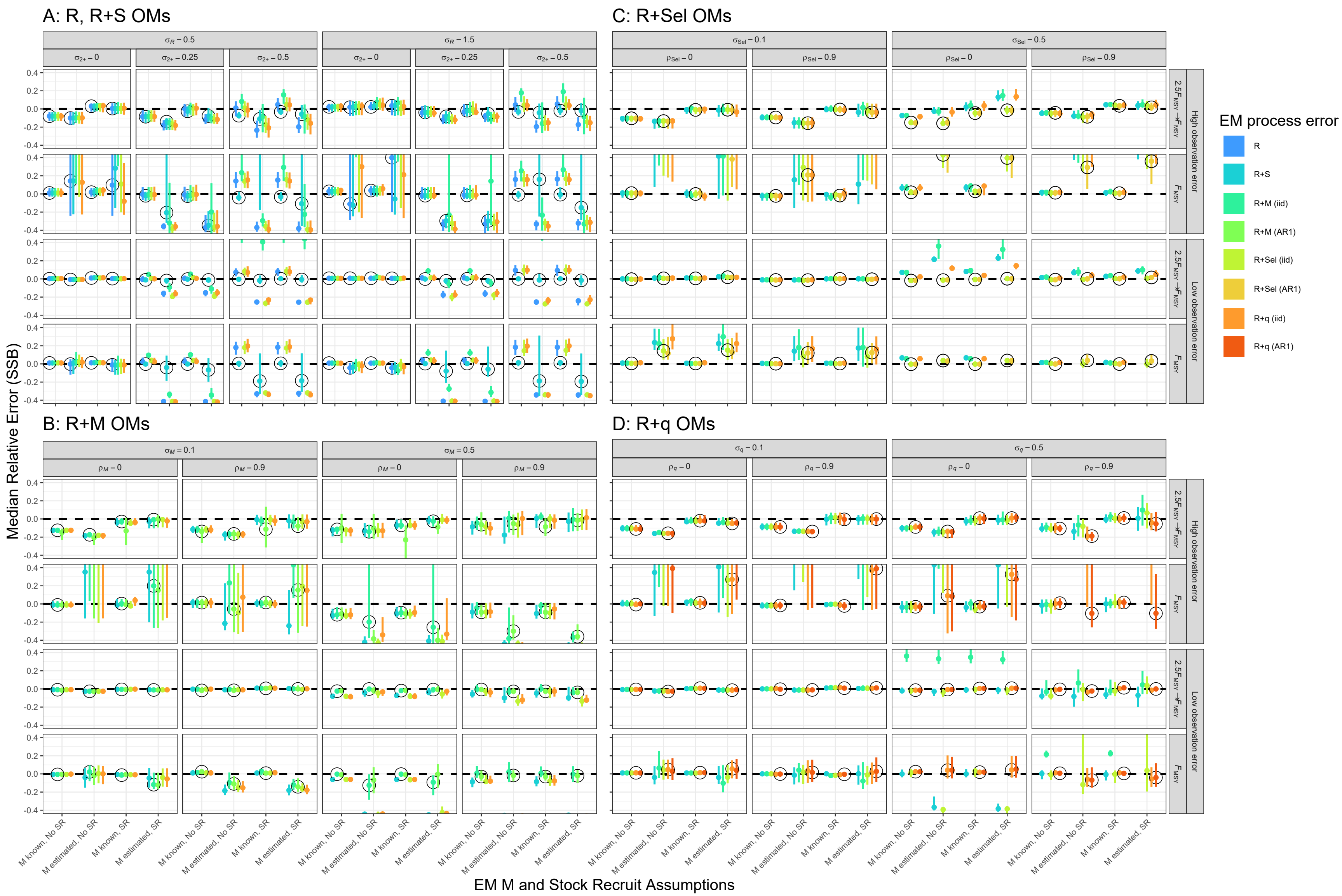

Fig. S6. Median relative error of terminal year SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and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

Fig. S7. Median relative error of 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.

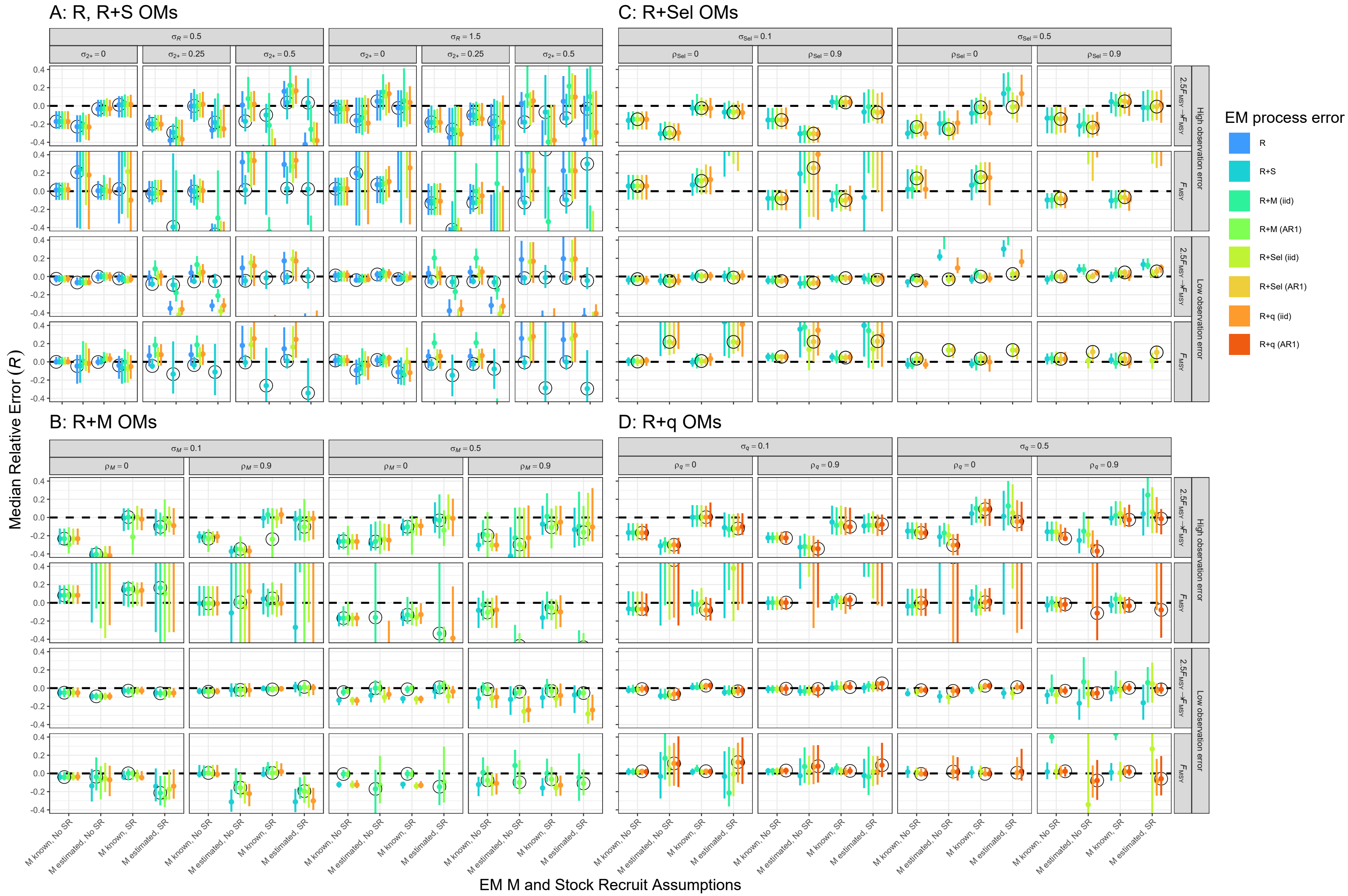

Fig. S8. Median relative error of terminal year recruitment for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating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

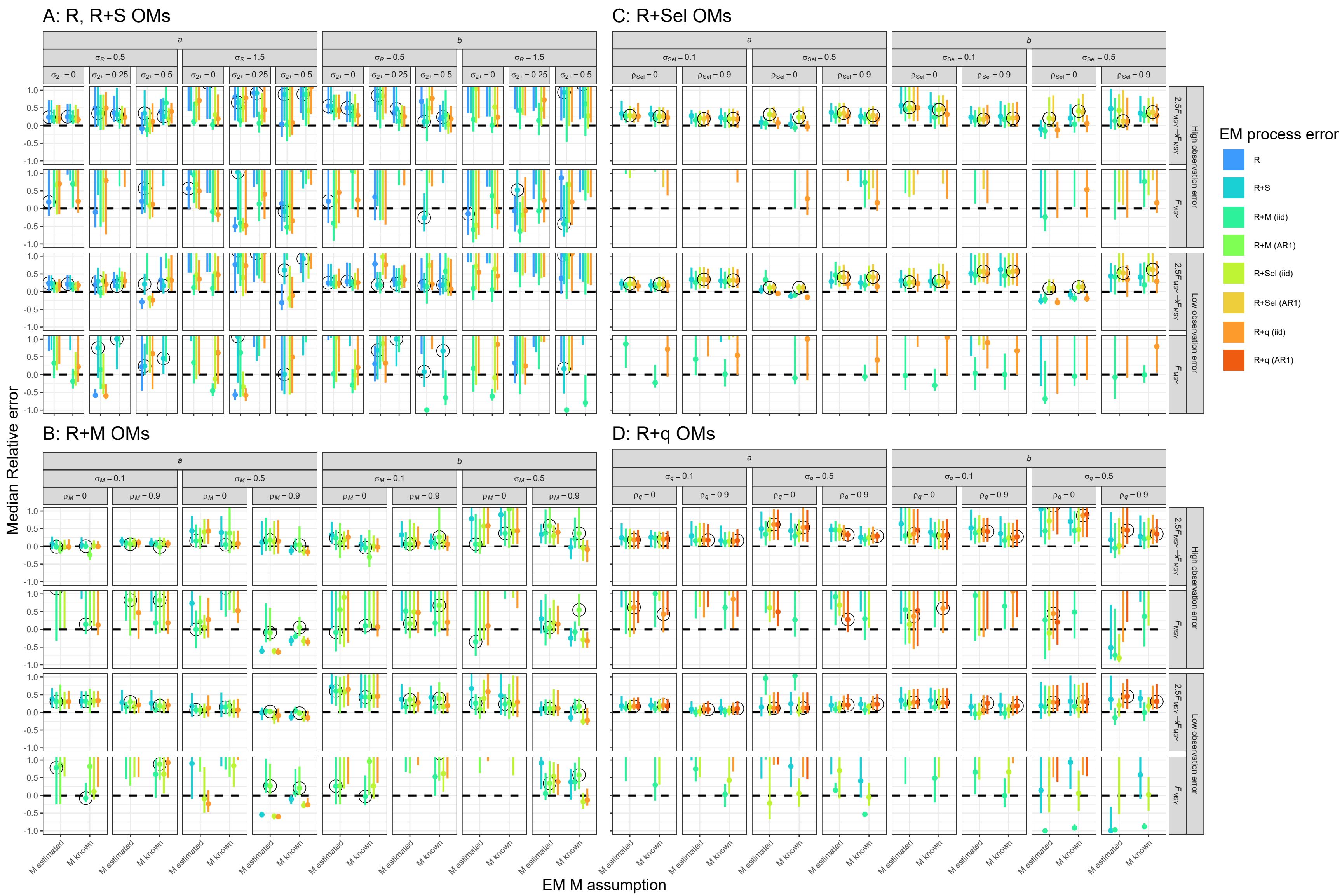

Fig. S9. Median relative error of Beverton-Holt stock-recruit parameters (a and b) for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

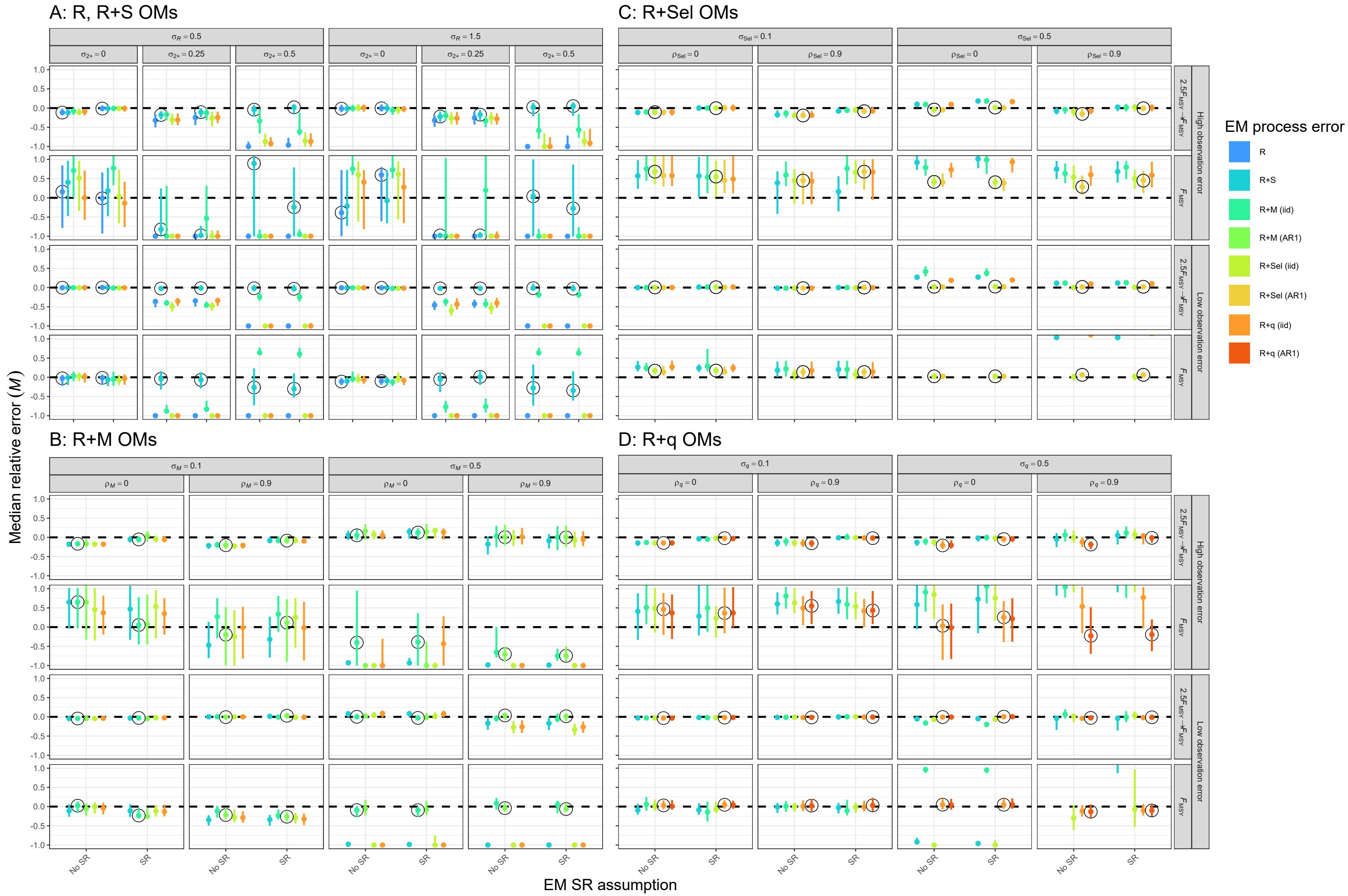

Fig. S10. Median relative error of median natural mortality for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating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

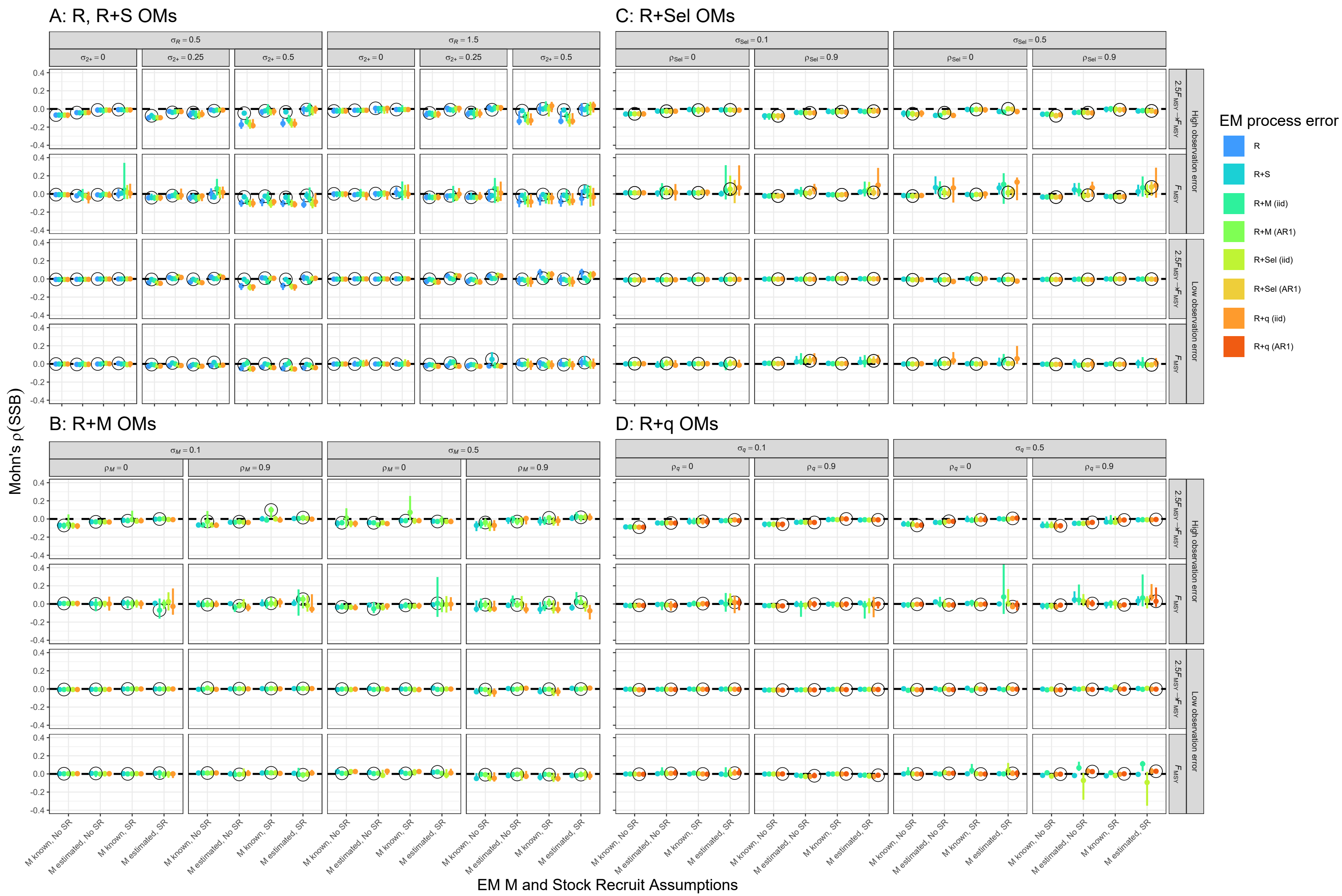

Fig. S11. Median Mohn's ρ for SSB for estimating models fitted to data sets simulated with alternative process error structures: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). Circled values indicate results where the EM process error structure matches that of the operating model and vertical lines represent 95% confidence intervals.

Fig. S12. 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. S13. 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.

Table S3. 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		$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_2	1.5		$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_3	0.5	0.25	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_4	1.5	0.25	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_5	0.5	0.50	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
NAA_6	1.5	0.50	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	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		$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{14}	1.5		$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{15}	0.5	0.25	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{16}	1.5	0.25	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{17}	0.5	0.50	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
NAA_{18}	1.5	0.50	$2.5F_{\mathrm{MSY}} \to F_{\mathrm{MSY}}$	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 S4. 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	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	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_{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	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	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_{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 S5. 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	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_2	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_3	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
Sel_4	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	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	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{10}	0.5	0.5	0.0	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{11}	0.5	0.1	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
Sel_{12}	0.5	0.5	0.9	$2.5F_{\text{MSY}} \rightarrow F_{\text{MSY}}$	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 S6. 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	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
q_2	0.5	0.5	0.0	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
q_3	0.5	0.1	0.9	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.1, Age composition SD = 0.3
q_4	0.5	0.5	0.9	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	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	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
q_{10}	0.5	0.5	0.0	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
q_{11}	0.5	0.1	0.9	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	Index SD = 0.4, Age composition SD = 1.5
q_{12}	0.5	0.5	0.9	$2.5F_{\mathrm{MSY}} \rightarrow F_{\mathrm{MSY}}$	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

Convergence full results

~~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.~~For many R and R+S OMs, convergence rate declined when either the median natural mortality rate or the Beverton-Holt SRR was estimated even when the process error assumptions of the EMs and OMs matched (Figure S14, A). When there was high observation error and constant fishing pressure ($F = F_{\text{MSY}}$ 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 S14, 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

Table S7. 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 ($\rho_{\text{Sel}} = 0$)	+	+	+	+
EM_{14}	Beverton-Holt	0.2	R+Sel ($\rho_{\text{Sel}} = 0$)	+	+	+	+
EM_{15}	Mean recruitment	Estimated	R+Sel ($\rho_{\text{Sel}} = 0$)	+	+	+	+
EM_{16}	Beverton-Holt	Estimated	R+Sel ($\rho_{\text{Sel}} = 0$)	+	+	+	+
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)	−	−	−	+

Table S8. For each OM process error type (columns), percent reduction in deviance for logistic regression models fit to indicators of convergence (maximum absolute gradient < 10^{-6}) with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM Process Error	30.40	0.45	17.57	16.04	24.03
EM M Assumption	2.38	24.11	4.42	1.02	2.66
EM SR Assumption	1.80	0.32	0.96	3.38	2.13
OM Obs. Error	0.12	0.77	0.33	1.76	0.28
OM F History	3.51	6.33	2.36	5.86	5.30
OM σ_R	<0.01	<0.01	–	–	–
OM σ_{2+}	–	<0.01	–	–	–
OM σ_M	–	–	0.39	–	–
OM ρ_R	–	–	0.09	–	–
OM σ_{Sel}	–	–	–	1.08	–
OM ρ_{Sel}	–	–	–	0.01	–
OM σ_q	–	–	–	–	0.06
OM ρ_q	–	–	–	–	<0.01
All factors	43.69	35.72	29.33	34.57	40.69
+ All Two Way	50.53	42.99	43.91	45.93	48.62
+ All Three Way	52.30	48.41	46.81	47.71	50.40

~~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.~~

Table S9. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to transformed Mohn's ρ values for each simulation (Eq. 3) for fishing mortality averaged over all age classes with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	0.06	0.09	0.01	0.12	0.01
EM SR assumption	0.01	<0.01	0.01	0.02	0.01
EM Process Error	0.03	0.07	0.02	0.06	0.03
OM Obs. Error	0.16	0.10	0.05	0.02	0.07
OM F History	0.07	0.02	0.03	0.24	0.03
OM σ_R	<0.01	0.01	–	–	–
OM σ_{2+}	–	0.09	–	–	–
OM σ_M	–	–	<0.01	–	–
OM ρ_R	–	–	<0.01	–	–
OM σ_{Sel}	–	–	–	0.01	–
OM ρ_{Sel}	–	–	–	<0.01	–
OM σ_q	–	–	–	–	<0.01
OM ρ_q	–	–	–	–	0.01
All factors	0.32	0.38	0.12	0.48	0.15
+ All Two Way	0.65	0.67	0.30	0.95	0.43
+ All Three Way	1.18	1.11	0.63	1.34	0.90

~~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.~~

Table S10. For each OM process error type (columns), percent reduction in deviance for linear regression models fit to transformed Mohn's ρ values for each simulation (Eq. 3) for recruitment with each OM and EM factor (rows) included individually, combined, and with second and third order interactions.

Factor	R	R+S	R+M	R+Sel	R+q
EM M Assumption	0.86	0.56	0.16	1.00	1.27
EM SR assumption	<0.01	0.02	0.01	0.01	0.01
EM Process Error	0.01	0.59	0.18	0.07	0.04
OM Obs. Error	0.34	0.01	0.08	0.24	0.27
OM F History	0.91	0.22	0.06	1.20	1.67
OM σ_R	<0.01	0.14	–	–	–
OM σ_{2+}	–	0.11	–	–	–
OM σ_M	–	–	0.01	–	–
OM ρ_R	–	–	<0.01	–	–
OM σ_{Sel}	–	–	–	0.01	–
OM ρ_{Sel}	–	–	–	0.01	–
OM σ_q	–	–	–	–	0.01
OM ρ_q	–	–	–	–	0.01
All factors	2.28	1.74	0.51	2.66	3.51
+ All Two Way	4.20	2.74	1.08	5.08	6.51
+ All Three Way	4.83	3.79	1.79	6.03	7.82

~~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.~~

OMs with lower selectivity process error variability or higher observation error (Figure S14, 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 (S14, 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 S3). 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.

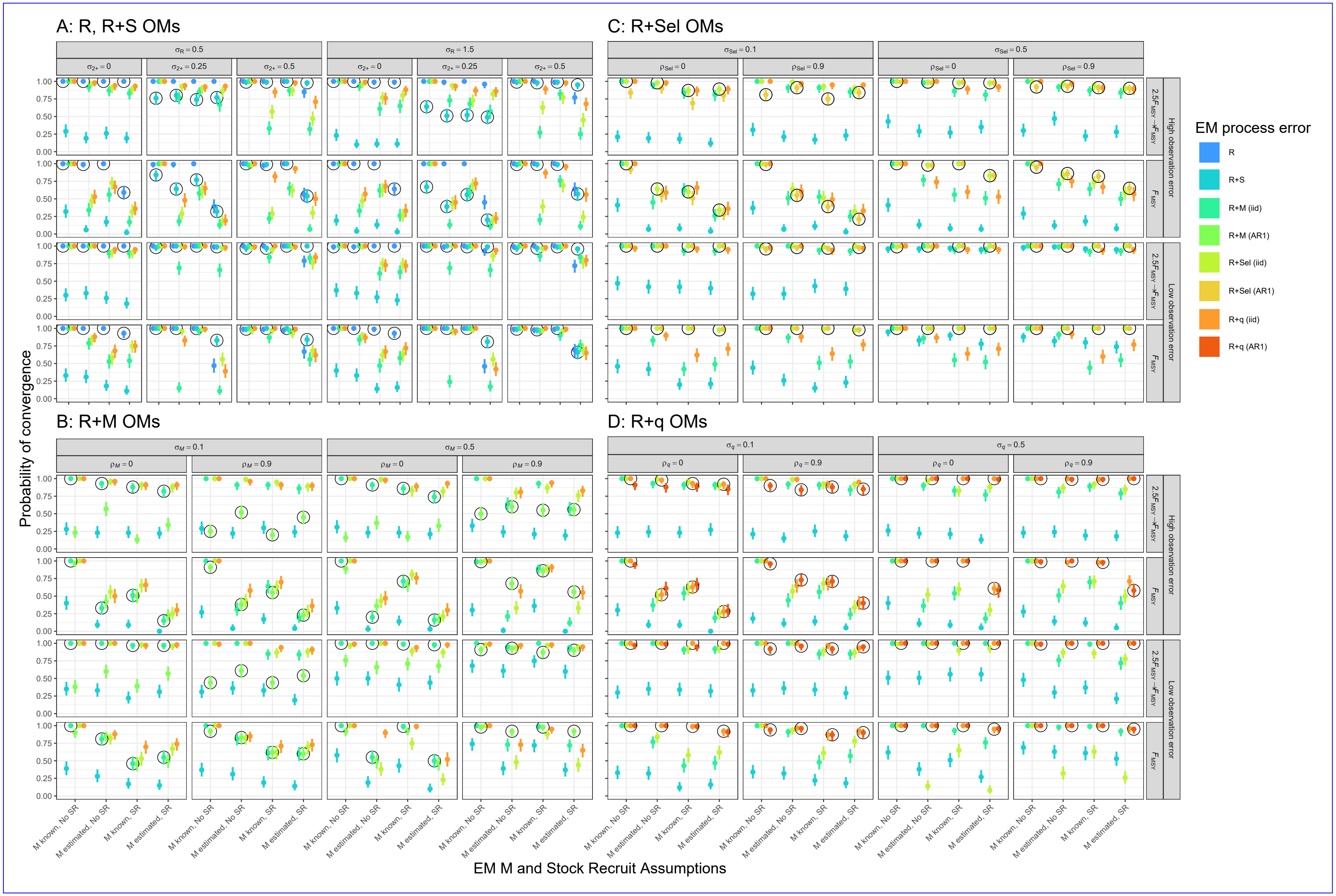

Fig. S14. ~~Median relative error~~ Estimated probability of ~~terminal year recruitment~~ fits providing hessian-based standard errors for ~~estimating models fitted to data sets simulated with~~ EMs assuming alternative process error ~~structures:~~ (colored points and lines), and median natural mortality (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. S15. ~~Median relative error~~ 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 ~~parameters~~ relationships (~~a and b~~estimated or not; along x-axis) ~~for estimating models~~ when fitted to ~~data sets simulated with alternative process error structures:~~ 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.

AIC for process error full results

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 S16, 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 S16, 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 S16, 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 S16,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.

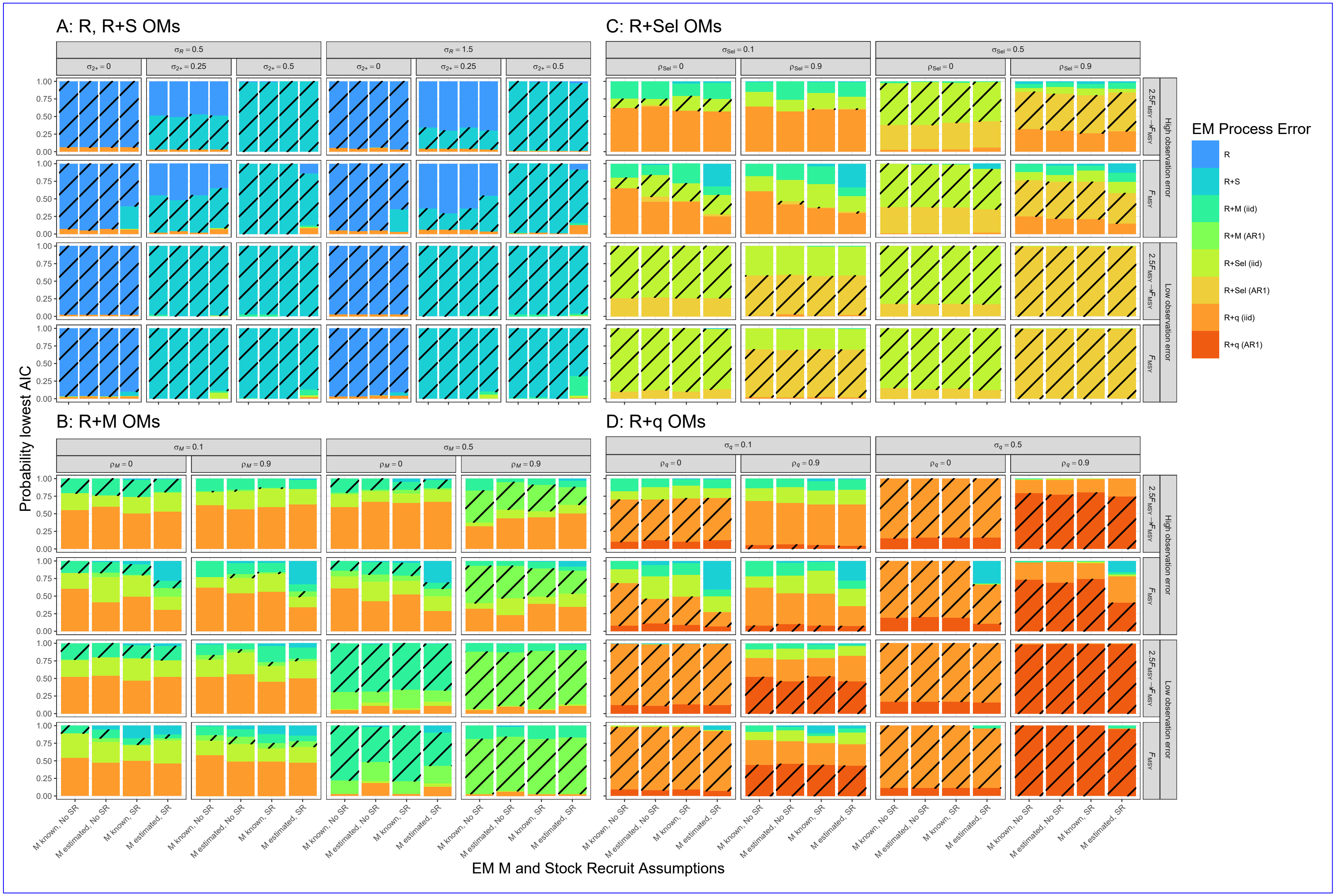

Fig. S16. ~~Median relative error~~ Estimated probability of ~~median natural mortality~~ lowest AIC for ~~estimating models fitted to data sets simulated with~~ EMs assuming alternative process error structures ~~:~~ (colored bars) conditional on alternative assumptions for median natural mortality (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. ~~Circled values~~ Striped bars indicate results where the EM process error structure matches that of the operating model~~and vertical lines represent 95% confidence intervals~~.

AIC for stock-recruit relationship full results

For each EM and OM combination, 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 recruitment would have larger variation in realized SSB.

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 S17).

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 S17, A). The largest variation in SSB occurred in OMs with larger recruitment variability ($\sigma_R = 1.5$; Figure S17, 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 S17, 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 S17, 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 S17, D). There was also little effect of variability or correlation of catchability process errors or the level of observation error on AIC accuracy.

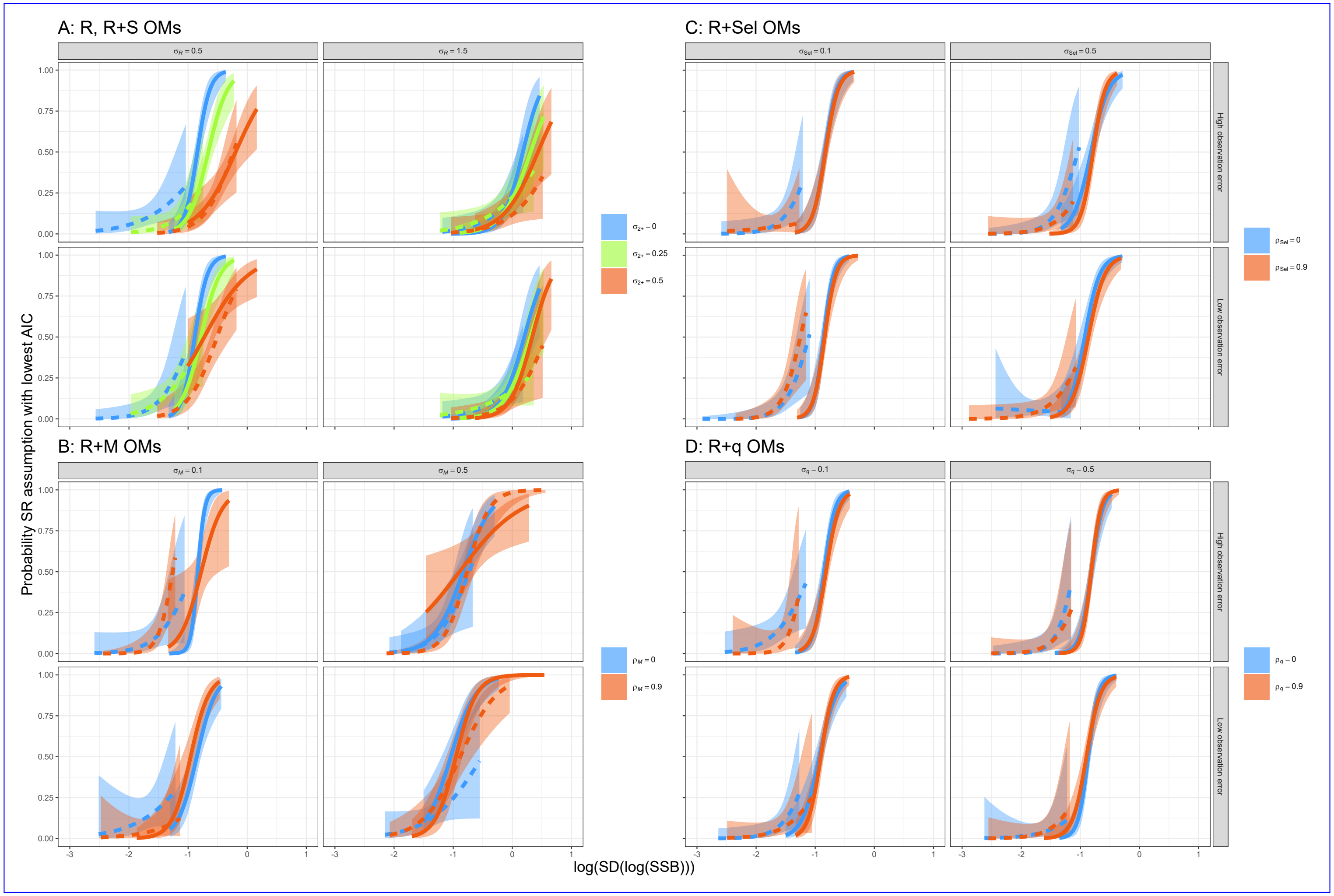

Fig. S17. ~~Median Mohn's ρ~~ Estimated probability of ~~fishing mortality averaged over all age classes~~ lowest AIC from logistic regression on the log-standard deviation of the true log(SSB) in each simulation for estimating ~~models fitted to data sets simulated~~ model with Beverton-Holt stock-recruit 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 ~~structures~~structure: R and R+S (A), R+Sel (B), R+M (C), or R+q (D)~~. Circled values indicate results where~~, and median M is assumed known in the EM~~process error structure matches that of the operating model~~. Solid and ~~vertical~~ 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.

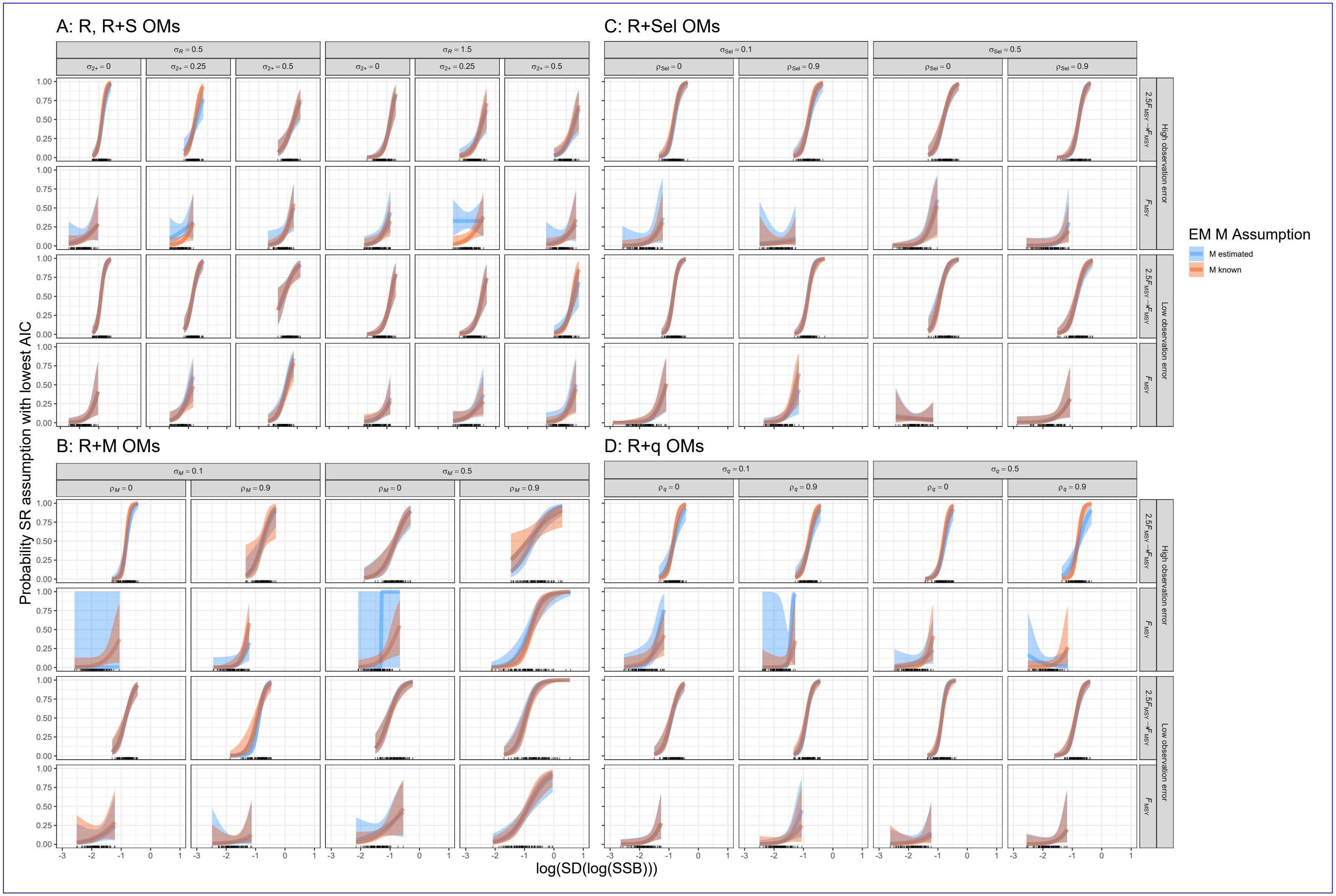

Fig. S18. ~~Median Mohn's ρ~~ Estimated probability of ~~recruitment~~ lowest AIC from logistic regression on the log-standard deviation of the true log(SSB) in each simulation for estimating ~~models fitted to data sets simulated~~ model with Beverton-Holt stock-recruit relationships, rather than the 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 ~~structures~~structure: R and R+S (A), R+Sel (B), R+M (C), or R+q (D). ~~Circled~~ Rug along x-axis denotes $SD(\log(SSB))$ values ~~indicate results where the EM process error structure matches that of the operating model~~ for each

Beverton-Holt parameter bias full results

Across all OMs, there was generally less bias and(or) lower variability in 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 S9, 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 S9, 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 S9, 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 had 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 S9, 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 S9, D, fourth row).

Median natural mortality rate bias full results

Across all OMs and EMs there was little effect of estimating SRRs on the bias in estimation of median natural mortality (Figure S10). 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 S10, 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 S10, 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 S10, 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 S10, 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.

Spawning stock biomass bias full results

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 S6, 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 S6, 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 S6, 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 S6, 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 ($\sigma_{\mathrm{Sel}} = 0.5$) or low selectivity process errors ($\sigma_{\mathrm{Sel}} = 0.1$) 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 S6, 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 S6, 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 S6, 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 S6, 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 S8, 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.

Mohn's ρ Full results

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 S11, 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 S11,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 S13, 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 S13, B-D).