Good practices, trade-offs, and precautions for model diagnostics in integrated stock assessments

Maia S. Kapur[1], Nicholas Ducharme-Barth[2], Megumi Oshima[2], and Felipe Carvalho[2].

[1]NOAA Fisheries, Alaska Fisheries Science Center, Seattle, WA, United States
[2]NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI, United States

Keywords: Good Practices; Fisheries Assessment; Integrated Analysis; Model Diagnostics

Corresponding author 1: Maia S. Kapur
email: maia.kapur@noaa.gov
phone: 206 526 4688

Abstract

Carvalho et al. (2021) provided a "cookbook" for implementing contemporary model diagnostics, which included convergence checks, examinations of fits to data, retrospective and hindcasting analyses, likelihood profiling, and model-free validation. However, it remains unclear whether these widely-used diagnostics exhibit consistent behavior in the presence of model misspecification, and whether there are trade-offs in diagnostic performance that the assessment community should consider. This illustrative study uses a statistical catch-at-age simulation framework to compare diagnostic performance across a spectrum of correctly specified and mis-specified assessment models that incorporate compositional, survey, and catch data. Results are used to contextualize how reliably common diagnostic tests perform given the degree and nature of known model issues, including parameter and model process misspecification, and combinations thereof, and trade-offs among model fits, prediction skill, and retrospective bias that analysts must consider as they evaluate diagnostic performance. A surprising number of mis-specified models were able to pass certain diagnostic tests, although there was a trend of more frequent failure with increased mis-specification for most diagnostic tests. Nearly all models that failed multiple tests were mis-specified, indicating the value of examining multiple diagnostics during model evaluation. Diagnostic performance was best (most sensitive) when recruitment variability was low and historical exploitation rates were high, likely due to the induction of better contrast in the data, particularly indices of abundance, under this scenario. These results suggest caution when using standalone diagnostic results as the basis for selecting a "best" assessment model, a set of models to include within an ensemble, or to inform model weighting. The discussion advises stock assessors to consider the interplay across multiple dynamics. Future work should evaluate how the resolution of the production function, quality and quantity of data time series, and exploitation history can influence diagnostic performance.

1. Introduction

Sustainable exploitation of renewable natural resources requires quantitative scientific guidance. Integrated population dynamics models (e.g., Fournier and Archibald (1982)) have flourished as the tool of choice to evaluate the status and possible future outcomes for exploited, threatened or managed populations (Maunder and Punt, 2014; Tempel et al., 2014). Integrated population models use mathematical relationships (processes) to specify how changes in population abundance occur and to link model predictions to data (observations). The processes are themselves governed by parameters that can either be estimated during the model fitting process or pre-specified based on independent studies. With the increase of computing power and the popularization of integrated stock assessment modeling (Maunder et al., 2009), the complexity of modern stock assessment modeling for fisheries management has increased. Analysts have multiple ways to model the observed data and underlying population processes when conducting a stock assessment, yet the tradeoffs among modeling choices are not always obvious. Developing and presenting multiple candidate models has become commonplace for many national and international fishery management agencies (Karp et al., 2022). Sometimes, analysts and/or review bodies must choose among candidate models, or systematically filter and combine models in an ensemble (Jardim et al., 2021).

The complexity of data types and model structures available to fishery stock assessment, and the desire for objective means of evaluating multiple candidate models, have led to a growing interest in diagnostic tests. Diagnostic tests can serve an important role in model validation, a crucial step in the assessment process that establishes the credibility and robustness of the advice that proceeds from an assessment model. Model validation communicates confidence in model outputs to stakeholders not directly involved in model construction, though validation of key derived quantities (current biomass and fishing mortality, for example) is not possible, as these values cannot be directly observed. Therefore, diagnostic tests used for model validation typically focus on evaluating how well the model fits to the observed data, whether the model meets its statistical assumptions and seems ecologically plausible, and the robustness of the model to new or removed data. These tests remain an integral component to the evaluation and provision of management advice, and are often required in assessment reports.

A large number of studies over the last decade have specifically investigated the reliability of such tests, particularly in the context of model mis-specification. Previous work in model diagnostics has evaluated only a single source of mis-specification in relatively simple models (e.g., Carvalho et al. (2017); Piner et al. (2011)). Model specification decisions are related to the functional form of the process, the variables they depend on (e.g., age or length, Lee et al. (2019)) and potential temporal variability in those processes. Inappropriate specification of a population dynamics model can occur in the observation (data) and/or population processes (Maunder and Piner, 2017). **These can include spatial variability, local depletion, movement dynamics, and the precision and accuracy of data inputs, among others.** In contrast to population processes, the parameters of the observation model are nearly always estimated because they address sampling uncertainties that are largely unknowable without an estimate of the population dynamics. Mis-specification occurs when a process is assumed to be governed by the wrong functional form, a parameter is set to the wrong value, or a process is modelled such that temporal variability is not correctly accounted for (or even ignored). Mis-specification of the population processes can lead to biased estimates of the parameters and hence quantities of management interest while mis-specification of observation processes can lead to the data not providing the correct information about the estimated parameters (e.g., Langseth et al., 2016; Maunder et al., 2023). Moreover, mis-specification in one process can lead to poor fits to data directly linked to that process and to data indirectly linked via the population dynamics because all data and model processes are linked through the population dynamics equations (Lee et al., 2019; Taylor et al., 2013). The linkage of all data via the population dynamics equations is the strength of integrated modelling, but this strength also makes locating mis-specifications challenging. Almost certainly, diagnosing and correcting model mis-specification becomes more difficult when multiple processes are simultaneously mis-specified.

Carvalho et al. (2017) presented a simulation-based exploration into how popular diagnostic tests respond to standalone misspecifications for a singular stock. The chief finding of that study was that the examined diagnostic tests are not equally reliable at detecting model mis-specification (Carvalho et al., 2017). Residual analyses appeared best at detecting mis-specification of the observation model, while only the age-structured production model (ASPM, Maunder and Piner, 2015) could detect a mis-specification of the system dynamics (Carvalho et al., 2017). No single diagnostic could realiably identify the process of a given misspecification for complex models (such as those with many fisheries and/or data types). The key limitations of that work were that the simulations 1) did not consider varied life history strategies, particularly those that result in highly stochastic population trajectories, 2) did not consider varied levels of fishing mortality, which can impact the degree of contrast in simulated or real datasets and therefore affect parameter estimability (Magnusson et al 2007), and 3) model misspecifications were explored individually, so synergistic effects of misspecifying multiple parameters and/or processes on diagnostic performance remain unexplored. Finally, the results were not contextualized alongside the relative error in management quantities, so it was impossible to compare the presence and degree of bias in the models that failed diagnostic tests, versus those that did not. These caveats are especially important given that diagnostic tests are employed across a diverse range of stocks, particularly those managed by international organizations (Karp et al., 2022), whose life and exploitation histories may vary considerably from the initial study, and the fact that most stock assessors manipulate several model components at once (Maunder and Punt 2014), presenting many opportunities for misspecification to be introduced or eliminated. All of these limitations are revisited in the present study.

Carvalho et al. (2021) provided a "cookbook" for implementing modern model diagnostics and model performance evaluation. That work suggested that a model would be considered adequate for management advice if the model a) optimizes successfully, b) fits to the data (e.g., passes a residual analysis), c) provides reliable estimates of population trends and scale, d) produces consistent results when provided new data, or if data are removed (e.g., retrospective analysis), and e) can make adequate future predictions (e.g., hindcasting). An ideal situation would allow for the suite of diagnostic tests presented in Carvalho et al. (2021) to be used to select among candidate models, or to evaluate or weight a set of candidate models for inclusion in an ensemble (aka Jardim et al., 2021). Such a one-size-fits-all approach is not yet realized, particularly since real-world applications of the cookbook have encountered tradeoffs between passing all or most diagnostics. Implementing the cookbook has also become complicated because some diagnostics do not have clear thresholds for passing, or if they do, the applicability of such thresholds to a diversity of stocks has not been rigorously tested. For example, since the publication of threshold-like values for rho (Hurtado-Ferro et al., 2012), many management agencies have made a practice of selecting among management models based on whether they fall among the cutoffs presented in that paper (Carvalho et al., 2021; Merino et al. 2022). Recent work has shown that these cutoffs should not be considered universal (e.g., Breivik et al. 2023), and proposed alternative approaches to model selection (e.g., the "Rose" approach, Legault, 2020). Much uncertainty remains about the appropriateness of strict thresholds for many diagnostic criteria, the order in which they should be applied, and how to consider models that perform well on some, but not all diagnostics.

This paper synthesizes the lessons learned from previous simulation work on diagnostic performance (Carvalho et al., 2017; Carvalho et al., 2021) and a series of workshops held with stock assessors (Karp et. al., 2022, Maunder et al., 2022) to explore and propose "good practices" for the application of diagnostics to integrated stock assessment models used for fisheries management. To contribute to the simulation-based literature on this topic, this paper presents an illustrative (but not exhaustive) study using a statistical catch-at-age simulation framework to compare diagnostic performance across a spectrum of correctly specified and mis-specified assessment models. The results are used to contextualize how reliably various diagnostic tests perform given the degree and nature of known mis-specifications in parameters and processes and trade-offs, among model fits, prediction skill, and retrospective bias that analysts must consider as they evaluate diagnostic performance.

The field of model diagnostics for fisheries assessment is emerging; the development of state-space modeling applications also warrant new diagnostic approaches as they become more commonly used in assessments (Li et al., 2024). The discussion includes good practices and an evaluation of the tradeoffs in diagnostic performance that analysts must consider when developing and selecting models used for fisheries management.

2. Methods

2.1. Overview

We use a combinatory simulation approach (Figure S3) to introduce a variety of misspecifications into the estimation method, and to evaluate the estimation performance diagnostic tests. The simulation procedure first involves the specification of a model of true population dynamics, the operating model (OM), using the R package *ss3sim* (Johnson et al., 2019). The OM is used to generate typical data for a fish population (time-series of catches in weights from a fishery fleet, an index of abundance from a survey, and the proportions-at-length and -age for both the fishery and surveys). All three types of data are generated from the OM for one hundred years, with a period of early recruitment deviations extending for 26 years prior to the start of the model. These generated data are used in a set of estimation methods (EMs). The EMs fit to the data and estimate the quantities of management interest. These estimates are then compared to the true values from the OM.

2.2. Operating model (OM)

The OM is an age-structured population dynamics model implemented in the Stock Synthesis software (SS version 3.30.16, Methot and Wetzel, 2013). Key systems and observation processes and their parameter values are listed in Table 1, and some biological assumptions (e.g., stock-recruitment steepness, natural mortality, and growth) were originally estimated for Pacific cod (*Gadus macrocephalus,* Anderson et al., 2014. A general description of the OM is as follows: it is a one-area, single-sex model, with time-invariant length-weight, length-at-age, and maturity-at-age relationships, and natural mortality (*M*). Recruitment is assumed to follow a time-invariant Beverton and Holt (1957) relationship with steepness (expected recruitment at 20% of the expected pre-fishery biomass, *h*) set to 0.65 and randomly-generated stock-recruitment deviation. The observation process involves a single fishing fleet and survey.

The relative probability of capture at length (selectivity) for the fishery fleet and survey is time-invariant; the length at 50% selectivity is 52 cm and 51 cm for the fishery and survey, respectively. All ages are available to the survey and fishery fleets. The initial conditions were specified so that there was no impact of fishing prior to the first year.

2.3. OM Scenarios

Six OM scenarios were designed using combinations of fishing mortality (*F*) vectors and various levels of recruitment variability: 0.1, 0.4 or 1.0 Simulations were designed to produce unbiased estimates of spawning biomass in the absence of misspecification (Figure 1a). For each of the six OM scenarios, 16 OM replicates were generated by resampling the data given a) process error, sampling a vector of recruitment deviates (Figures 1b and S1) from a normal distribution with mean zero and the applicable variance for that scenario, and b) observation error for the catch, survey, and compositional datasets (Figure 1c, d, and e), described below. This number of replicates was chosen to compromise between simulation run times and balance among the randomized experimental design. Two vectors of annual fishing mortality rates were generated: 1) a "high" fishing mortality scenario increases to a maximum of twice the true F_{MSY} and then decreases to $0.9F_{MSY}$, and 2) a "low" fishing mortality scenario is defined by the overall maximum fishing mortality rate equal to one-fourth of natural mortality (Figure S1). Process error in each OM replicate arises from variability in annual recruitment deviations (Figure 1c) see Reproduction section below) and fishing mortality time series. Sixteen replicates of the OM were generated for each of the six possible fishing mortality and recruitment variation combinations (e.g., low fishing mortality and recruitment variance of 0.4, Figure S1), for a total of 96 unique OMs.

2.4. Data Generation

Data used in the EMs are the time-series of catches in weight from the fishing fleet, a time series of relative abundance from the survey, and length- and age-composition data that provide a measure of the size and age structure of the survey and the fishery (Figure S2). The catch observations were assumed to be known without error (coefficient of variation = 0.01). Each abundance observation was assumed to be proportional to the available absolute abundance, called "catchability" in fisheries applications, and was generated from a log-normal distribution with a coefficient of variation of 0.2 (Figure 1b). Each length- and age-composition observation was generated from a multinomial distribution with variability described by an effective sample size of 50 (Figure 1d,e). No additional data weighting was applied to any component. Below, we describe the model components that were manipulated in our simulation experiments and how the misspecifications were implemented. The order of the corrections varied with each simulation, such that all possible unique combinations of corrections were explored.

2.5. Mis-specified processes

Growth

The growth curve in the OM is modeled using the von Bertalanffy (1957) growth function, a common relationship used in fisheries assessment to model the length (cm) of an average fish with respect to its age (years). The model is parameterized using asymptotic length (the inferred length at infinite age) and the growth rate (the rate at which the average fish reaches asymptotic length). Researchers may obtain inaccurate input values of this parameter via unrepresentative or imprecise sampling, which fails to capture or correctly measure individuals at large lengths and/or older ages (Shelton & Mangel 2012). In the correctly-specified estimation method all main growth parameters (*L1*, L_∞, *K*, and CVs of length at ages) were freely estimated. For estimation methods exploring a mis-specification in growth, *L1* and *K* were set to correct values and L_∞ was set to the mis-specified value, while CVs of length at ages remained estimable. Estimation methods with L_∞ mis-specified are denoted by the letter L.

Natural mortality

Generally, it is difficult to obtain empirical estimates of natural mortality for any fish species (e.g., Hamel, 2014; Punt et al., 2021; Maunder et al., 2023). In fisheries, several methods infer this value from the maximum age or length (Then et al., 2015 and Hamel and Cope, 2022) or via a meta-analysis of similar species within a genus (Thorson et al., 2017). In the OM, natural mortality, *M*, is time- and age-invariant and set at 0.2 yr^{-1} (Table 1). Estimation methods with natural mortality mis-specified at incorrect values are denoted by the letter M.

Reproduction

Steepness of the stock-recruitment relationship, the common measure of stock resilience, is a highly uncertain yet critical quantity in fishery stock assessment and management. Estimating steepness inside a stock assessment model is difficult and estimates have lower precision and higher bias (Lee et al., 2012). Because of the difficulty of estimating steepness, this parameter is typically not estimated within assessment models. Annual reproduction *R* in the OM is calculated based on a Beverton-Holt function (Equation 1) of the system-wide reproductive biomass in a given year (*SB*), expected unfished recruitment R_0 and biomass SB_0 and *h*, i.e.:

$$R_y = \frac{4hR_0SB_y}{SB_0(1-h)+SB_y(5h-1)}e^{-0.5\sigma_R^2+\tilde{R}_y};\ \tilde{R}_y \sim N(0,\sigma_R^2) \qquad \text{Eq. 1}$$

Annual recruitment deviates, governed by a recruitment variability error term (σ_R^2), measure the log-distance from the deterministic curve given in Equation 1 and is a source of process error in the OM. The variance in recruitment deviates was set to either 0.1, 0.4 or 1.0. Recruitment deviates are randomly generated once for each OM replicate; steepness and R_0 are not estimated. In the estimation methods, the recruitment deviates and R_0 are estimated with steepness set to either the correct or a mis-specified value (Table S1); σ_R^2 is set to the correct value from the applicable OM. Estimation methods with steepness mis-specified are denoted by the letter H.

Fishery selectivity

In the OM, the fishery and survey have a length-based double normal selectivity pattern with the initial selectivity at first bin and final selectivity at last bin parameters set to low numbers to avoid numerical estimation. This creates an asymptotic selectivity curve, meaning that all individuals above a certain size have a close to equal probability of being captured. When the selectivity is correct, estimation methods estimate the selectivity parameters under the assumption that selectivity is an asymptotic function of length for the fishery. Estimation methods with selectivity mis-specified are denoted by the letter X, indicating that the model sets the ascending limb (e.g., the length at 50% selectivity) to a mis-specified value (Table S1).

Determining misspecification thresholds

Instead of arbitrarily choosing mis-specified parameter values, the two values nearest to those used in the OM that led to a 10% change in the final-year depletion (the ratio between final-year biomass and expected unfished biomass) were solved for. The threshold detection was performed by fitting a series of estimation methods to the same dataset generated by the OM across a broad range of fixed values for each parameter in turn: for example, 19 estimation methods with steepness *h* set to 0.05, 0.10, 0.15,…,0.95 and other parameter values estimated. This was done for each unique combination of recruitment deviation variance and fishing mortality. The relative error in final-year depletion was calculated between the estimates from each estimation method and the values in the OM and used to find the parameter values nearest to the OM values that corresponded to relative errors of -10% and 10% (Table S1). Preliminary investigations included OM values that resulted in relative errors of as much as 20%, but these often involved most parameters hitting their bounds; ±10% was selected to keep most parameters within their plausible ranges (and to avoid having to discard models where estimates were on bounds). This step ensured that the mis-specifications implemented in the experimental design are known to impact estimated outputs to the same extent. This led to two mis-specified parameter values, one above and one below the values used in the OM, for all parameters except for steepness. In cases where no values above the OM value met the mis-specification threshold criteria (e.g. steepness), the value above the OM value that corresponded to the greatest relative error was selected.

2.6. Estimation methods and experimental design

The EMs were implemented in Stock Synthesis version 3.30 (Methot and Wetzel, 2013). The experimental design followed a systematic procedure (Figure 2), which enabled the determination of how well model diagnostics could detect the nature and extent of model misspecification. Calculation of model diagnostics across 1,536 EMs was facilitated by using the OpenScienceGrid HTCondor high-throughput computing network (Pordes et al., 2007; Sfiligoi et al., 2009) and the *ssgrid* package in R (Ducharme-Barth, 2022). The experimental workflow was as follows:

1. Generate an operating model "replicate" with process errors (recruitment deviations and fishing mortalities) and observation errors (generation of survey abundance indices and compositional data).
2. Sample a vector of four values for each replicate, each with an even probability of being either a 0 or 1. This vector determines how each mis-specification, H, M, X, or L, is implemented. A value of 0 indicates the mis-specification is below the true value whereas

a value of 1 indicates the mis-specification is above the true value. For example, the first OM replicate may have the draw [0, 0, 0, 0] in which all four parameters would be specified below the true value for all EMs fit to those OM data. The next OM replicate may have a different vector draw, ensuring that variation caused by differences in process and observation errors are balanced against the directionality of mis-specifications.

3. Fit EMs for each of the 16 functionally unique combinations corresponding to the mis-specified categories (Table S1) to each replicate. All unique combinations of mis-specifications were evaluated. For example the combination “HMXL” denotes a model with all four mis-specifications, while “MX” denotes a model with only natural mortality (M) and selectivity (X) mis-specified. Note that “MX” is functionally equivalent to “XM” so only the former is investigated. EMs using all components correctly specified and using the correctly-stratified data from the corresponding OM replicate are labeled as “correct”.
4. Repeat steps for each of sixteen resampled OM replicates. This protocol ensures the effect of the mis-specifications was not influenced by the high/low nature of the random vector assigned to each combination. In total, the study design fit 1,536 EMs (16 unique estimation methods fitted to 96 OM replicates).

2.7. Performance metrics

Relative Error

The results were summarized by the deviation between the estimates of the management quantities and the corresponding OM values. In lieu of fisheries-specific management quantities (e.g., the ratio of current biomass to the biomass that corresponds to maximum sustainable yield), we examined values common across the EMs, namely the time series in reproductive biomass (here, spawning stock biomass, *SSB*) and reproductive output (here, recruitment). In addition to the general trend in these estimated values, we also evaluated results based on the mean *SSB* over the last ten years. Together, these statistics aim to capture temporal variation in estimation performance as well as model performance during the recent period, which is typically of more interest to managers. The deviations between EM and OM values by year *y*, replicate *i*, combination *j* and scenario *k* were summarized using relative (equation 2) or absolute (equation 3) relative errors and then averaged across replicates.

$$\boldsymbol{MRE_{SSB_y}} = \sum_{i} \frac{\hat{S}SB_y^{EM_{ijk}} - SSB_y^{OM}i}{SSB_y^{OM}i} \boldsymbol{/i} \qquad \text{Eq. 2}$$

Both measures indicate the magnitude of difference between estimated quantities and the OM values. Relative error (positive or negative) enables us to investigate whether there are systematic and/or directional biases induced by the various mis-specifications. Using absolute relative error disregards the direction of the difference, and is useful for highlighting the scale of the effects of various mis-specifications. The mean absolute relative (MARE) errors for the terminal ten years of *SSB* are calculated via:

$$\boldsymbol{MARE}_{\boldsymbol{SSB_y}} = \sum_{i} \frac{\left| \sum_{y=91}^{100} \frac{\widehat{SSB}_y^{EM_{i,j,k}}}{10} - \sum_{y=91}^{100} \frac{SSB_y^{OM_i}}{10} \right|}{\sum_{y=91}^{100} \frac{SSB_y^{OM_i}}{10}} / \boldsymbol{i} \qquad \text{Eq. 3}$$

2.8. Model Diagnostics: Review and Application to Simulations

We applied model diagnostics following the recommendations of the cookbook using the associated R package *ss3diags* (Carvalho et al., 2021). The following section provides a brief summary of the logic and method behind each diagnostic, and how it was applied to our simulations.

Convergence

Models were assumed to have converged if no parameters were estimated at a bound, the gradient was relatively small (less than 1e-4) and the Hessian matrix was invertible, as recommended in Carvalho et al. (2021). The results shown here are comprised of converged models only.

Residual analysis

We explored non-random variation in residual patterns using a non-parametric runs test, wherein the 2-sided p-value is calculated for the distribution of residuals about a model estimate (typically estimated indices of survey abundance). If this p value is greater than or equal to 0.05, there is no evidence to reject the hypothesis that the residuals are randomly residuals and the model is determined to pass the runs test. We also calculated the root mean square error (RMSE) for the survey and compositional time series data as a measure of the standard deviation of the residuals from the model estimates. A small RMSE (≤ 0.3) indicates a reasonably precise model fit to relative abundance indices (Winker et al., 2018).

R_0 likelihood profile

We constructed likelihood profiles on unfished recruitment (*R_0*) using the `profile()` function from R package *r4ss* (Taylor et al., 2011). This approach sequentially fixes unfished recruitment at a pre-specified value and re-runs the estimation method with whatever other parameter settings were specified in the original experiment. This was repeated for all of the unique OM replicate-estimation method combinations described previously. The range of R_0 values used were chosen for each OM replicate, to encompass one unit of *R_0* (in log space) both above and below the MLE for the correct estimation method associated with that replicate, in increments of **0.2.**

This profile enables evaluation of the stability of the parameter estimate, which is influential in terms of model scale, and the relative influence of individual data sources upon the parameter. A poorly-estimated parameter is revealed by a profile that is flat (**delta** likelihood values below ~1.96 across a large parameter range), and/or may be characterized by data conflicts (where one or more data sources achieves a minimum likelihood at a much higher or lower parameter value than the others, or than the total likelihood. Wang et al. (2017) proposed the “psi” statistic, which quantifies whether the maximum likelihood estimate of R_0 for a specific data component falls within the 95% confidence interval for the total likelihood. This method has the potential to measure of the information content of a given likelihood component (lower values indicate less information, and are a rough measure of the degree of mismatch between the total likelihood for a given EM and the profile obtained for that data component). We did not implement the psi statistic in this study, as it is not widely used and the comparison of psi statistics from models with dissimilar parameterizations was not clear. Indications of poor parameter estimation or data conflict suggest that either model assumptions or data inputs need be re-evaluated.

Retrospective analysis

A retrospective analysis is a useful approach for addressing the consistency of terminal-year estimates. The analysis sequentially removes a year of data (a peel) at a time and reruns the model. The typical interpretation of this analysis is that serial over- or under-estimation of quantities such as SSB or fishing mortality are indicative of unidentified process error, and require a revisitation of model assumptions. The severity of over- or under-estimation is normally evaluated by eye and by the calculation of rho (equation 4) which is then compared to pre-determined thresholds (Hurtado-Ferro et al., 2015). We conducted retrospective analyses using the retro function from *r4ss* and mean rho over five, one-year peels was calculated as:

$$rho = \frac{1}{h}\sum_{t=1}^{h}\left(\frac{X_{T-t} - \widehat{\boldsymbol{X_{T-t}}}}{\widehat{X_{T-t}}}\right) \qquad \text{Eq. 4}$$

where X is the SSB or fishing mortality, $\hat{X}$ is the corresponding estimate from the reference model (model fitted to the full dataset), T is the terminal year of the model, and h is the number of peels (Hurtado-Ferro et al., 2015). Models with rho values less than -0.15 **or** greater than 0.20 would fail the retrospective diagnostic based on the rule of thumb as proposed by Hurtado-Ferro et al. (2015).

Age-structured production model

Maunder and Piner (2015) proposed an age-structured production model (ASPM) as a model diagnostic for complex age-structured integrated assessments. Briefly, this approach fixes selectivity, assumes average recruitment, and disregards compositional data. The tool can be used to determine whether the stock dynamics are readily explained by the production function and catches alone, which would suggest that the survey time series provides information regarding absolute abundance (Minte-Vera et al., 2017). A discrepancy between the ASPM and age-structured stock trajectory might indicate mis-specification of the components which make up the production function.

The ASPM performs best in situations characterized by high and low periods of fishing effort (also known as “contrast”) and where observations (i.e., catch, life history, and index) are reasonable representations of the actual states. It has been shown **to be sensitive to** misspecification of key systems-modeled processes that control the shape of the production function (Carvalho et al., 2017). However, failure of the ASPM is not necessarily indicative of model mis-specification and could be due to several factors. The stock could be recruitment driven (e.g., short-lived fishes with high recruitment variability) and/or lightly exploited such that the fishing signal is not strong enough to drive change in the stock.

A deterministic recruitment model is a similar means to diagnose a model's ability to capture the production function. Deterministic recruitment model is a simpler alternative to the ASPM as it only requires recruitment to be constrained to what would be predicted by the stock-recruit relationship without deviation (Merino et al. 2022). For both the ASPM and deterministic recruitment model, we calculated the relative difference in model estimates of R_0, MSY, and the mean absolute difference (MARE) in predicted SSB between the full model and the ASPM/deterministic recruitment model; these metrics are taken to measure how well-defined and influential the production function is upon stock dynamics, given the mis-specifications investigated in our study.

Hindcast cross-validation (MASE)

The accuracy and precision of a model's prediction skill can be measured with hindcast cross-validation, which involves comparing observations to predicted future values (Kell et al., 2022).. It is similar to retrospective analysis in that it involves peeling one year of data away at a time and re-fitting the model but involves an extra step of predicting the removed observation. The predicted values are cross-validated by comparing the model's one-step-ahead forecast, or expected value, of the observation at time t ($\underline{y_t}$) versus a "naïve" forecasted value equal to the last observation (y_{t-1}) for a given number of hindcasting time steps (h). The prediction skill can be calculated using the mean absolute scaled error (MASE) between models, where values less than one indicate that the model did better than the naïve approach:

$$MASE = \frac{\frac{1}{h}\sum_{\boldsymbol{t=T-h+1}}^{\boldsymbol{T}}\left|\underline{\boldsymbol{y_t}} - \boldsymbol{y_t}\right|}{\frac{1}{h}\sum_{t=T-h+1}^{T}|y_t - y_{t-1}|} \qquad \text{Eq. 5}$$

MASE was calculated for relative abundance indices and composition data using ten hindcast steps.

Recruitment trend

The principle of the goodness-of-fit tests (runs and RMSE) described above is that residual patterns in model estimates can be indicative of model mis-specification and un-modeled process. The estimation of recruitment deviates is a principle way that process error is incorporated into stock assessments, and residual trends therein may similarly indicate a mis-specification (uncaptured process error). Following Merino et al. (2022) the existence of a significant linear trend in the recruitment deviates was quantified, and monotonic trends, and non-monotonic (any) trends in recruitment deviates were tested. Additionally, it was calculated if first-order autocorrelation in the deviates was different from 0 and runs tests (using a threshold of $p \geq 0.05$ to pass) were applied to test for non-randomness.

3. Results

3.1. Convergence

Overall, 82% (1264/1536) of the models converged; results are only presented for converged models (

Table 2). All correctly-specified EMs converged, and convergence frequency across all scenarios declined as the number of mis-specifications increased to a minimum of 79% for four misspecifications. Proportionally fewer mis-specified models converged when fishing mortality was high, regardless of recruitment variability. These were typically disqualified due to gradients above the threshold. Models with low recruitment variability and low fishing mortality converged the most frequently overall, though convergence rates declined with increasing mis-specifications.

3.2. Relative error

The magnitude of error in estimated SSB and depletion varied among OM replicates, with systematic changes in error given by the fishing mortality vector and level of recruitment variability. The MRE of terminal SSB was highest with greater model-misspecification and greater variation in recruitment for both exploitation (*F*) scenarios, though the absolute value of error was greater under the high *F* scenario. This same pattern was present for MRE of estimated depletion, though the overall scale of error was smaller (ranging from -50 to 50%, Figure 1a).

3.3. Residual analysis

The residual analyses examined fits to the survey abundance time-series, as well as the calculation of the root-mean-square error (RMSE) for the survey abundance time-series, length, and age composition data. RMSE for fishery and survey compositional data increased systematically with an increasing number of mis-specifications, while the average RMSE for the survey index of abundance did not dramatically increase even in the presence of 3 or 4 mis-specifications (Figure 2a). Scenarios with high fishing mortality and/or low recruitment variability exhibited the largest increases in RMSE for composition data as the number of mis-specifications increased. Importantly, no models resulted in RMSE values above the 30% threshold indicated in Winker et al. (2018).

The majority (97%) of correct models passed the runs test; while pass rates declined with increasing numbers of misspecification, the overall failure rate only ranged from 4%-10% (Figure 3 and Table 2). Visual inspection of models that failed the runs test showed slightly worse fits to the data for the highly mis-specified models compared to the correct model (Figure 3). There were similarities between the performance of the runs test and RMSE. Firstly, most diagnostic responsiveness (e.g., increased failure rates with increased degree of misspecification) emerged for the compositional data while p-values for the survey index of abundance were less responsive (Figure 3). There also appeared to be greater sensitivity (more failures) to increased mis-specification when fishing mortality was high and/or recruitment variability was low, as was seen for RMSE.

3.4. R_0 likelihood profile

Of the 254,064 unique models run as part of the profiling exercise, 99% converged and were included in this analysis. Figure 4 presents the likelihood profiles for the total objective function, survey data and length and age composition data, scaled so that the x-axis represents the difference between the fixed R_0 for the model at hand and the value for R_0 from the OM. The MLE for R_0, indicated by the total likelihood, was well-defined for correct EMs. For EMs with zero or one misspecifications, the total likelihood agreed with the survey and length-composition data, while the age composition data indicated R_0 values slightly lower than the other data sources. The likelihood profiles differed systematically from those obtained using the correct EM upon the introduction of two or more mis-specifications. The qualitative and relative behavior of the profiles was strikingly consistent within EMs: the survey data were always the broadest, and the length and age composition profiles were consistently narrower than the survey and shifted slightly below the total MLEs. Profiles for the age composition data had lower specificity overall (many statistically indistinguishable models above and below the MLE). Conflict between the best R_0 values of the survey and length composition data versus the age composition data was present in all EMs.

3.5. Retrospective Analysis

The thresholds proposed by Hurtado-Ferro et al. (2015) had little ability to detect model mis-specification in our framework. Overall, 5% of the EMs had rho values for the spawning biomass and fishing mortality time series outside of the thresholds [-0.15, 0.2] (Figure 5). For correct EMs at all levels of fishing mortality and recruitment variability, rho values for both *SSB* and fishing mortality were centered around 0 with very few models falling outside the thresholds, though higher rho values occurred with the highest level of recruitment variability explored. For the low *F* scenarios, rho values for SSB and fishing mortality did diverge from zero to a greater degree than the high *F* scenarios. The change in rho was most pronounced for scenarios with low F and high recruitment variability. EMs with high fishing mortality did not show clear trends in the magnitude or direction of rho values with increasing levels of misspecification nor across recruitment variability levels.

3.6. Age-structured production model

The performance of the ASPM varied by scenario. Generally, the ASPM estimated SSB trajectories that were higher in scale and smoother through time when fishing mortality was low (Figure 6). The ASPM was better able to capture the scale and dynamics of the SSB trajectory under the high *F* scenario, with minimal difference from the base model under high *F* and low recruitment variability (Figure 6). The ASPM and model with deterministic recruitment consistently resulted in lower MSY.

Both the ASPM and deterministic recruitment results showed virtually identical patterns in terms of relative error in R_0 from the full model, and the MARE of *SSB* (Figure 7). Both the ASPM and deterministic recruitment were able to estimate R_0 well. However, the MARE of *SSB* was consistently over estimated. There were differences between the trends of MARE of *SSB* across the number of mis-specifications between the ASPM and deterministic recruitment model. For the ASPM models, MARE of *SSB* showed a general increase as the number of mis-specifications increased across all levels of fishing mortality and recruitment variability. For the deterministic recruitment models at all recruitment variability levels, MARE of *SSB* decreased as the number of mis-specifications increased for models with low fishing mortality but increased as the number of mis-specifications increased for models with high fishing mortality Both ASPM and deterministic recruitment models with high recruitment variability had the smallest difference in MSY from the full model and models with low recruitment variability had the greatest difference in MSY from the full model.

3.7. Hindcast Cross Validation (MASE)

The MASE statistic indicated that models had better predictive performance than the null (e.g. MASE < 1) for all levels of misspecification, with increasing predictive performance with fewer misspecifications (Figure 8). However, the proportion of models that passed this diagnostic only ranged from 54% (fully misspecified) to 66% (correct model). There were not strong patterns in MASE statistics across data types nor *F* levels, though it seemed that the lowest passing rates occurred under lower levels of recruitment variability (54% at the lowest, to 68% under the highest values of sigmaR). Of the models that had worse predictive power than a null model (MASE > 1), the failed statistic most commonly occurred for age-composition data (Figure 8). Overall, the hindcast statistic had the highest failure rates of all diagnostics regardless of exploitation level or recruitment variability (Table 2).

3.8. Recruitment trend

Significant linear trends in recruitment deviates usually indicated the presence of at least one model mis-specification at least for low fishing mortality scenarios, and increasing the number of mis-specifications tended to increase the proportion of model runs that showed significant linear trends in recruitment deviates (Figure 9). However, a substantial proportion of mis-specified models did not indicate significant trends in recruitment deviates (false negatives), and some correctly specified models showed significant linear trends in the recruitment deviates (false positives). Additionally, rates of false positives and false negatives were not consistent across OM replicates. Testing for the presence of monotonic trends or any (non-monotonic) trend showed similar results as the test for linear trends in recruitment deviates. Runs tests of the recruitment deviates and testing for any non-zero first order autocorrelation indicated a poor ability to discriminate between correctly specified models and mis-specified models under the low fishing mortality scenario.

4. Discussion

4.1. Limitations

Data Richness

Several characteristics of our study design limit the interpretation of our results and form the basis for future research regarding the utility and robustness of diagnostic tools. The simulations explored here are centered on a data-rich, age-structured assessment model, with a longer time series of data (particularly compositional data) than is likely available even for the most heavily-monitored stocks (Maunder et al., 2014; Ono et al., 2015). The experiment presented here was designed to eliminate data concerns so that the performance of diagnostics tests could be evaluated in a "best-case scenario"; we did not wish to construct candidate EMs that were so mis-specified that they would be dismissed out of hand by any competent analyst (e.g., an extinct population, survey estimates completely out of range). A potential risk of our simulation design is that these data are *so* informative and abundant that models are able to approximate the correct solution (i.e., fit the survey time series to a satisfactory degree) even when parameters are mis-specified. This could explain the apparent lack of power that the diagnostics appear to have for discerning between correctly specified and mis-specified models, particularly for the survey time-series. Model diagnostics that did not perform well in our study are unlikely to perform well for similar stocks with fewer or worse data. An additional research avenue related to this topic is the diagnostic use of changes to the effective sample size for compositional data under a Dirichlet-multinomial (D-M) distribution (Thorson et al., 2023). However, a minority of global models have reliable compositional data to begin with, and a minority of those use the D-M distribution in estimation routines. This highlights the primacy of developing and testing diagnostic tools that are applicable to a range of model types.

A simulation that explores how diagnostic performance varies with a reduction in time series length of frequency, smaller compositional sample sizes, or larger observational errors would test this hypothesis. Relatedly, we did not introduce temporal variability into our simulation framework, which may have dampened our ability to detect a retrospective trend. Model-specific confidence intervals can be calculated for rho (Miller & Legault, 2017), though this approach has not been adapted widely. We suggest further research into the topic of retrospective thresholds; recent work has indicated that retrospective performance indeed varies across model complexity and the amount of data provided to the model (Breivik et al., 2023) or the breadth of model configurations considered in an ensemble (Brooks & Brodziak, 2024). Our results suggest that even in the absence of temporal variability, the combination of low exploitation levels, high recruitment variability and/or high levels of model misspecification can produce patterning in rho values, so it is not inconceivable that thresholds specific to recruitment and exploitation histories could be developed.

Data Quality

Most assessments, particularly those that rely on fisheries-dependent data sources, will utilize data that are biased to an unknown degree. This study assumes that all data used in the EMs are representative and unbiased relative to the dynamics of the OM, again a deliberate decision to represent a 'best-case scenario'. In addition to the data availability issues discussed above, the impact of data *quality* on diagnostic performance remains an open question for future research (Punt 2023, this issue; Liljestrand et al., 2024). Processes such as hyperstability, effort creep, or the under-reporting of catch can result in non-proportional indices of relative abundance. The presence of these dynamics could manifest through the residual runs test, poor ASPM, hindcast cross-validation, or by inducing a trend in the recruitment deviates. It is possible that some diagnostics are more useful for identifying *data* mis-specifications rather than parameter or model mis-specifications. An urgent area of future research is to investigate the performance of diagnostic tests in models with well-specified processes and parameters but poorly-representative data.

Recruitment Driven Dynamics and Model Parsimony

This study is also limited because the operating model appears to be recruitment-driven, meaning that the biomass dynamics suggested by the age-structured model are distinct from what the underlying production function would suggest, so the recruitment time series (and deviations thereof) explain the stock's trajectory. (In contrast, a "production driven" stock would be one where the time series of biomass is well-explained by the mean stock-recruitment relationship and historical fishery removals). This is likely because of the "data-richness" of the simulation, in that the composition data (from which recruitment estimates are derived) was abundant and continuous throughout the time series. The varied performance of the ASPM diagnostic across scenarios is consistent with findings that such tools perform best when applied to production, not recruitment-driven stocks (Minte-Vera et al, 2017). This relates to the above discussion of model and data complexity, and is important considering that many global stocks do not consider age structure at all for management purposes. We emphasize that several tools, particularly goodness-of-fit tests and explorations of model convergence are applicable across a range of model types.

All operating models tested here used one of two vectors for fishing mortality. Some diagnostics, like MASE, might simply echo the stock's responsiveness to fishing pressure, which in this case will be more pronounced in trajectories that have lower *SSB* because of reduced recruitment. As stated in Punt et al. (2023, this issue), process error can occur in multiple model processes, including selectivity. This study does not investigate the impacts of time-varying selectivity curves, or allowing the estimation of the descending limb of the double normal curve, which could enable the model to compensate for additional mis-specified processes. However, given that many mis-specified models were able to pass various diagnostic tests, we anticipate that introducing further flexibility into the model structures would reinforce the ability for mis-specified models to satisfy diagnostic criteria. An investigation into the relative performance of these production-related diagnostics for stocks with and without well-informed production functions would be informative.

4.2. How do individual diagnostics perform?

Fits to the Data and Parameter Estimation

Model convergence was the strongest indicator of the number of misspecifications, which is consistent with our recommendation (and that of the Cookbook) that it be the first diagnostic test performed, and alternative structures explored if the test is failed. Out of all diagnostics other than model convergence, the RMSE test most reliably returned higher (poorer) values with an increasing amount of misspecifications. This is reassuring evidence that goodness-of-fit tests can be a useful first step in evaluating a model. In contrast, the runs test using the traditional cutoff of 0.05 was one of the least reliable diagnostics. A majority of highly mis-specified models passed the runs test at this threshold, corresponding to the fact that all models seemed to visually fit the survey index. Given this result, it is possible that the statistical cutoff for passing the test is not appropriate. Our results suggest that correct models have, on average, p-values of 0.5 or higher – though the range was uninformatively large (<0.05 to 0.95). It is also illustrative that diagnostic performance was most robust (i.e., failure rates higher with increased mis-specification) in scenarios with high fishing mortality and/or low recruitment variability. This corresponds to previous studies that have indicated that model contrast is often required to inform stock dynamics (Magnusson and Hilborn, 2007), and suggests caution for analysts applying goodness-of-fit tests to lightly exploited stocks. Similarly, both residual diagnostics exhibited greater variation when applied to compositional data, while survey time-series scores remained flat. This result is likely related to the production question described above, and underscores the importance of running diagnostics on multiple data sources (when available). Analysts should consider whether visually satisfactory fits to survey abundance time series are sufficient for model acceptance, and be warned that there are circumstances (high observation errors, abundance of other data sources) that can lead a mis-specified model to fit the survey time series well.

Likelihood profiles appear to remain internally consistent, with the relative degree of conflict stable across mis-specified models (e.g. survey abundance data were always less informative than fishery and survey length composition data, with broader profiles more distinct from the total likelihood). This indicates that likelihood profiles can remain a useful tool for determining data conflicts and information content regardless of the degree of misspecification in an assessment model, but would not alert the analyst to the presence of misspecification.

Model Consistency

The present study complements recent research to offer new insights into the process of identifying and addressing retrospective patterns in fish stock assessments. Legault (2020) compared the rho-adjustment to the “Rose” approach, a time-intensive process whereby a retrospective pattern is eliminated across an ensemble of models, allowing the analyst to change multiple processes or data inputs. That evaluation determined that both approaches are viable for removing retrospective patterns, though neither identifies the cause(s), and the choice between approaches depends on the time and expertise available. rho calculated from retrospective analyses in our study was a surprisingly poor correlate to model misspecification (given the traditional cutoff range of -0.15 to 0.2 for SSB, Hurtado-Ferro et al. 2015), though retrospective performance did degrade with increasing mis-specifications. This does not indicate that the retrospective diagnostic is a poor tool, rather that the presence of a retrospective pattern (and associated failure of the rho cutoff) is not a guaranteed outcome when the parameters we examined are mis-specified. This finding is similar to those of Breivik et al. (2023) who indicated that the acceptable range for the traditional rho diagnostic varies with the amount of data and type of model used. The authors proposed an alternative “post-sample rho significance test” with the aim to reduce subjectivity in decisions about significant retrospective patterns in state-space assessment models. The new statistic, which conditions the distribution of rho values on the data prior to the retrospective period, enables the analyst to evaluate whether the retrospective pattern is truly anomalous or reasonable given the model used. This presents a promising avenue for future research, though readers are reminded that retrospective patterns can be reduced while reference points remain biased (Szuwalski et al. 2018). We do not propose alternative, universal ranges for the original rho statistic.

Prediction skill

For most data components, MASE scores were related to the number of model misspecifications, suggesting that the hindcast diagnostic can evaluate model performance and can potentially detect model misspecification. However, the MASE criterion used here to quantify prediction skill was not sensitive enough to detect mis-specification across all models; this is consistent with earlier work indicating that good MASE performance for hindcasting is likely to occur if the stock is production driven and the production function is estimable from the data (Minte-Vera et al., 2021), which is not the case in our example. It is notable that the MASE statistic was the most commonly failed across all levels of F and recruitment variability (Table 2), and that the high F – low recruitment scenario did not exhibit improved performance of this diagnostic as it did for other tests. Given that MASE tests explicitly for prediction skill, it's understandable that scores less than 1 are harder to obtain when the precision of the data is high and the biomass trend is relatively flat; this is the reason why users may elect to set a precision threshold for the naïve prediction error below which the statistic is no longer penalized, an area that requires further research.

The ASPM and deterministic recruitment model diagnostics appear versatile and promising. Carvalho et al., (2017) showed via simulation analysis that ASPM was the only diagnostic capable of detecting mis-specification of the key systems-modeled processes that control the shape of the production function. Here, the ASPM and the deterministic recruitment model were not able to provide evidence for a production function, essentially confirming this example as a recruitment-driven model. The deterministic recruitment model returned virtually the same results as the ASPM, so either diagnostic could be used as an alternative to measure the effects of fishing. Our findings suggest that the ASPM performs best under scenarios with low-to-medium recruitment variability and when fishing mortality is high, such that contrast is induced in the time series – an emergent theme among both estimation and diagnostic performance. These findings, as well as those for the MASE diagnostic, underscore the importance of data contrast and, relatedly, the presence of a production function in determining diagnostic performance, which was

Estimated recruitments are one of the primary ways process error is modeled in stock assessments, so examining the recruitment deviates for trend and non-randomness makes intuitive sense as a potential model diagnostic (Merino et al. 2022). Merino et al. (2022) explored using a test for statistical significance of a linear trend in the recruitment residuals as a potential diagnostic for identifying model mis-specification within an ensemble of models. Our study is the first time (to our knowledge) that this diagnostic has been formally evaluated within a simulation framework. As currently formulated, this diagnostic may have some discriminatory power in identifying mis-specified models from correctly specified models. While it was more likely that models with significant linear trend in the residuals were mis-specified, there was still a chance (~7%) that the model was correctly specified (false-positive). This is close to the assumed false positive rate of the statistical test ($p<=0.05$). However, there remains a large false negative rate for mis-specified models. Further simulation testing is needed to refine either the statistical thresholds used to identify significant residual trend to see if that improves discriminatory power or the types of mis-specifications this test may be used to identify.

4.3. Good practices in applying model diagnostics

Updating the "Cookbook"

The original cookbook (Carvalho et al., 2021) proposed a linear workflow of diagnostic tools, whereby a model is required to "pass" a set of diagnostics in a given order. That workflow inherently prioritized certain tests such as residual diagnostics and the runs test before likelihood profiling or retrospective analyses. The spirit of that approach – that a model should converge and reasonably fit the data to be considered a candidate – is unchanged, and we point readers to both the original workflow as well as the ordered list provided in Table 2 when applying diagnostic tests. Our findings further contextualize how modelers should use the outcomes of diagnostics: firstly, it is evident that the degree of recruitment variability and exploitation history together modulate diagnostic performance (likely through the induction of contrast in the data), so quantitative thresholds for most diagnostics, if desired, would need to be developed with those factors in mind. Secondly, diagnostics of prediction skill (namely the MASE statistic) appear less insensitive to model misspecification overall, though with more promising performance for age-composition data than for survey biomass. Further research into the best way to test for and improve prediction skill, particularly the use of thresholds or minima for such diagnostics, is warranted. Finally, our results show that it is possible to develop plausible, realistic stock assessments that fit data well and still perform poorly on some model diagnostics. We suggest that the community should strive for a balance among the considerations of model realism and diagnostic performance.

Tradeoffs in Model Development

The primary challenge in developing diagnostic workflows arises because stock assessors must evaluate a small subset of total possible models representing a population, far fewer than the hundreds of thousands of models run for this simulation analysis. In our study, the RMSE, ASPM and likelihood profile diagnostics were the most internally consistent and responsive to the presence of misspecification. Yet an assessment scientist would only see results for, at most, a dozen models, and have no knowledge of how divergent the selected model is from reality. Furthermore, the information gleaned from diagnostics such as the RMSE is not much more useful than a simple visual inspection of the model fits; it is likely that models with poor RMSE scores would have been discarded in the first place based on their poor fits to the survey data. This means that the scientist's holistic evaluation of the model's ecological plausibility remains necessary.

The tentative "good practices" and associated precautions presented in Table 3 warrant a comment about the general push towards automation of assessment procedures. We assert that stock assessment modeling requires the experience and the subjective evaluation of competing priorities, which are not replaceable by a set of diagnostic algorithms – particularly when the true recruitment trend might be unknown, as discussed above. Tools such as machine learning (particularly for image classification), boosted regression trees, and artificial intelligence present a promising avenue that may improve data collection (Zhang et al., 2022) and detect patterns in population dynamics (Mendoza et al, 2012; Memarzadeh et al., 2019). Furthermore, the nature of assessment science requires analysts to place value on sometimes competing priorities, whether in a formal framework such as a management strategy evaluation (Punt et al., 2016), or in the process of data or model weighting (Francis, 2017). These subjective tasks invite the consideration of socio-economic topics and the participation of fishery stakeholders, which could lead to model configurations being selected despite poor performance on one or more diagnostic criteria. For this reason, as well as the growing body of evidence that standardized cutoffs for diagnostic performance are not ideal for the selection of management models, we discourage the use of automatic pass/fail criteria for most diagnostic tests. Instead, analysts are encouraged to couple the results of diagnostic tests with their expert evaluation of the model's plausibility, given the biological and historical context of the stock, and in consultation with managers. We caution assessors against reverting to simplified assessment types (e.g., data-limited methods, Legault et al. 2023) in order to pass diagnostic tests as they carry the risk of poor management performance.

4.1. Conclusion

This study substantially expands the simulation framework developed by Carvalho et al. (2017) and updates the framework for applying diagnostics to integrated fisheries assessments presented by Carvalho et al. (2021). There remains several outstanding research avenues as the community continues to refine (or discard) quantitative diagnostic criteria. Further investigation of diagnostic performance should evaluate 1) the impact of changes to data quality or availability (particularly for the case of data-limited stocks (e.g., no age or length structure); 2) if there are correlations between the operating model characteristics (e.g. general stock trajectory or production/recruitment driven dynamics, or trends in process error) and diagnostic performance, and 3) whether the introduction of time-varying components (such as recruitment regime shifts, or time blocks in selectivity) impact the performance of diagnostic tests, particularly those associated with prediction skill. The scientific assessment community should continue to investigation via simulation without neglecting the subjective expertise and decision-making skill required to produce analyses for scientific management.

5. Acknowledgements

The authors would like to thank the participation of numerous scientists in the diagnostic workshops held over the last two years, and for the detailed discussions held with Henning Winker. We thank Melissa Haltuch for a careful review of the first version of this manuscript.

6. References

Anderson, S.C., Monnahan, C.C., Johnson, K.F., Ono, K., Valero, J.L., 2014. ss3sim: An R Package for Fisheries Stock Assessment Simulation with Stock Synthesis. PLOS ONE 9, e92725. https://doi.org/10.1371/journal.pone.0092725

Beverton, R. J. H. and S. J. Holt. On the Dynamics of Exploited Fish Populations. Fisheries Investment Series 2. 19. U. K. Min. of Agr. and Fish. Chapman and Hall: London (1957).

Breivik, O.N., Aldrin, M., Fuglebakk, E., Nielsen, A., 2023. Detecting significant retrospective patterns in state space fish stock assessment. Can. J. Fish. Aquat. Sci. 80, 1509–1518. https://doi.org/10.1139/cjfas-2022-0250

Brooks, E.N., Brodziak, J.K.T., 2024. Simulation testing performance of ensemble models when catch data are underreported. ICES Journal of Marine Science fsae067. https://doi.org/10.1093/icesjms/fsae067

Carvalho, F., Punt, A.E., Chang, Y.-J., Maunder, M.N., Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fisheries Research. 192, 28–40. https://doi.org/10.1016/j.fishres.2016.09.018.

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. https://doi.org/10.1016/j.fishres.2021.105959

Ducharme-Barth, N. 2022. ssgrid: ssgrid: Stock Sythesis - OpenScienceGrid - utilities. https://github.com/N-DucharmeBarth-NOAA/ssgrid, https://n-ducharmebarth-noaa.github.io/ssgrid/.

Fisch, N., Shertzer, K., Camp, E., Maunder, M., 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 fsad138. https://doi.org/10.1093/icesjms/fsad138

Francis, R.I.C.C., 2017. Revisiting data weighting in fisheries stock assessment models. Fisheries Research 192, 5–15. https://doi.org/10.1016/j.fishres.2016.06.006

Fournier, D., Archibald, C.P., 1982. A General Theory for Analyzing Catch at Age Data. Canadian Journal of Fisheries and Aquatic Sciences 39, 1195–1207. https://doi.org/10.1139/f82-157

Hamel, O.S., 2015. A method for calculating a meta-analytical prior for the natural mortality rate using multiple life history correlates. ICES Journal of Marine Science 72, 62–69. https://doi.org/10.1093/icesjms/fsu131

Hamel, O.S., Cope, J.M., 2022. Development and considerations for application of a longevity-based prior for the natural mortality rate. Fisheries Research 256, 106477. https://doi.org/10.1016/j.fishres.2022.106477

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., Punt, A.E., 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science 72, 99–110. https://doi.org/10.1093/icesjms/fsu198

Ian G. Taylor, Kathryn L. Doering, Kelli F. Johnson, Chantel R. Wetzel, Ian J. Stewart, 2021. Beyond visualizing catch-at-age models: Lessons learned from the r4ss package about software to support stock assessments, Fisheries Research, 239:105924. https://doi.org/10.1016/j.fishres.2021.105924.

Jardim, E., Azevedo, M., Brodziak, J., Brooks, E.N., Johnson, K.F., Klibansky, N., Millar, C.P., Minto, C., Mosqueira, I., Nash, R.D.M., Vasilakopoulos, P., Wells, B.K., 2021. Operationalizing ensemble models for scientific advice to fisheries management. ICES Journal of Marine Science 78, 1209–1216. https://doi.org/10.1093/icesjms/fsab010

Johnson KF, Anderson SC, Doering K, Monnahan CC, Stawitz CC, Taylor IG (2019). ss3sim: Fisheries Stock Assessment Simulation Testing with Stock Synthesis. R package version 1.0.3.

Karp, M.A., Kuriyama, P., Blackhart, K., Brodziak, J., Carvalho, F., Curti, K., Dick, E.J., Hanselman, D., Ianelli, J., Sagarese, S., Shertzer, K., Taylor, I., 2022. Common model diagnostics for fish stock assessments in the United States (No. NMFS-F/SPO240A). National Marine Fisheries Service.

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: is it me or my model talking? ICES Journal of Marine Science 78, 2244–2255. https://doi.org/10.1093/icesjms/fsab104
Langseth, B.J., Schueller, A.M., Shertzer, K.W., Craig, J.K., Smith, J.W., 2016. Management implications of temporally and spatially varying catchability for the Gulf of Mexico menhaden fishery. Fisheries Research 181, 186–197. https://doi.org/10.1016/j.fishres.2016.04.013

Li, C., Deroba, J.J., Miller, T.J., Legault, C.M., 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. https://doi.org/10.1016/j.fishres.2024.106968

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., Muffley, B., 2023. Data-rich but model-resistant: an evaluation of data-limited methods to manage fisheries with failed age-based stock assessments. Can. J. Fish. Aquat. Sci. 80, 27–42. https://doi.org/10.1139/cjfas-2022-0045

Legault, C.M., 2020. Rose vs. rho: a comparison of two approaches to address retrospective patterns in stock assessments. ICES Journal of Marine Science 77, 3016–3030. https://doi.org/10.1093/icesjms/fsaa184

Lee, H.-H., Maunder, M.N., Piner, K.R., Methot, R.D., 2012. Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models? Fisheries Research 125–126, 254–261. https://doi.org/10.1016/j.fishres.2012.03.001

Lee, H., Piner, K.R., Taylor, I.G., Kitakado, T., 2019. On the use of conditional age at length data as a likelihood component in integrated population dynamics models. Fisheries Research 216, 204–211. https://doi.org/10.1016/j.fishres.2019.04.007

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., Muffley, B., 2023. Data-rich but model-resistant: an evaluation of data-limited methods to manage fisheries with failed age-based stock assessments. Can. J. Fish. Aquat. Sci. 80, 27–42. https://doi.org/10.1139/cjfas-2022-0045

Liljestrand, E.M., Bence, J.R., 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. https://doi.org/10.1016/j.fishres.2024.107023

Magnusson, A. and Hilborn, R. (2007), What makes fisheries data informative?. Fish and Fisheries, 8: 337-358. https://doi.org/10.1111/j.1467-2979.2007.00258.x

Maunder, M.N., Schnute, J.T., Ianelli, J. 2009. Computers in fisheries population dynamics. In: Megrey, B.A., Moksness, E. (Eds.), Computers in Fisheries Research. Springer, pp. 337–372.

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fisheries Research 192, 16–27. https://doi.org/10.1016/j.fishres.2016.04.022

Maunder, M.N., Punt, A.A., 2014. A review of integrated analysis in fisheries stock assessment. Fisheries Research 142, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025

Maunder, M.N. and Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES Journal of Marine Science, 72(1), pp.7-18.

Maunder, M., Punt, A., Carvalho, F., Winker, H., Valero, J., Minte-Vera, C., 2022. 1st Workshop On Improving The Risk Analysis For Tropical Tunas In The Eastern Pacific Ocean: Model Diagnostics In Integrated Stock Assessments 1st Workshop On Improving The Risk Analysis For Tropical Tunas In The Eastern Pacific Ocean: Model Diagnostics In Integrated Stock Assessments (No. WSRSK-01).

Maunder, M.N., Hamel, O.S., Lee, H.-H., Piner, K.R., Cope, J.M., Punt, A.E., Ianelli, J.N., Castillo-Jordán, C., Kapur, M.S., Methot, R.D., 2023. A review of estimation methods for natural mortality and their performance in the context of fishery stock assessment. Fisheries Research 257, 106489. https://doi.org/10.1016/j.fishres.2022.106489

Memarzadeh, M., Britten, G.L., Worm, B., Boettiger, C., 2019. Rebuilding global fisheries under uncertainty. Proceedings of the National Academy of Sciences 116, 15985–15990. https://doi.org/10.1073/pnas.1902657116

Mendoza, M., Pennino, M.G., Bellido, J.M., 2012. Tree-based machine learning analysis for fisheries research. Fishery Management 61–75.

Merino, G., Urtizberea, A., Fu. D., Winker, H., Cardinale, M., Lauretta, M., Murua, H., Kitakado, T., Arrizabalaga, H., Scott, H., Pilling, G., Minte-Vera, C., Xu, H., Laborda, A., Erauskin-Extramiana, M., Santiago, J. 2022. Investigating trends in process error as a diagnostic for integrated fisheries stock assessments. Fisheries Research. https://doi.org/10.1016/j.fishres.2022.106478.

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

Timothy J. Miller, Christopher M. Legault, Statistical behavior of retrospective patterns and their effects on estimation of stock and harvest status, Fisheries Research, Volume 186, Part 1, 2017, Pages 109-120, ISSN 0165-7836, https://doi.org/10.1016/j.fishres.2016.08.002.

Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., Satoh, K., Uosaki, K., 2017. Get the biology right, or use size-composition data at your own risk. Fisheries Research. 192, 114–125. https://doi.org/10.1016/j.fishres.2017.01.014.

Minte-Vera, C.V., Maunder, M.N., Aires-da-Silva, A.M., 2021. Auxiliary diagnostic analyses used to detect model misspecification and highlight potential solutions in stock assessments: application to yellowfin tuna in the eastern Pacific Ocean. ICES Journal of Marine Science 78, 3521–3537. https://doi.org/10.1093/icesjms/fsab213

Ono, K., Licandeo, R., Muradian, M.L., Cunningham, C.J., Anderson, S.C., Hurtado-Ferro, F., Johnson, K.F., McGilliard, C.R., Monnahan, C.C., Szuwalski, C.S., Valero, J.L., Vert-Pre, K.A., Whitten, A.R., Punt, A.E., 2015. The importance of length and age composition data in statistical age-structured models for marine species. ICES Journal of Marine Science 72, 31–43. https://doi.org/10.1093/icesjms/fsu007

Piner, K.R., Lee, H.-H., Maunder, M.N., Methot, R.D., 2011. A Simulation-Based Method to Determine Model Misspecification: Examples Using Natural Mortality and Population Dynamics Models. Marine and Coastal Fisheries 3, 336–343. https://doi.org/10.1080/19425120.2011.611005

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy A., Avery, P., Blackburn, K., Wenaus, T., Würthwein, F., Foster, I., Gardner, R., Wilde, M., Blatecky, A., McGee, J., Quick, R. 2007. The open science grid. Journal of Physics: Conference Series. 78:12057. doi: 10.1088/1742-6596/78/1/012057

Punt, A.E., Butterworth, D.S., de Moor, C.L., De Oliveira, J.A.A., Haddon, M., 2016. Management strategy evaluation: Best practices. Fish and Fisheries 17, 303–334. https://doi.org/10.1111/faf.12104

Punt, A.E., Castillo-Jordán, C., Hamel, O.S., Cope, J.M., Maunder, M.N. and Ianelli, J.N., 2021. Consequences of error in natural mortality and its estimation in stock assessment models. Fisheries Research, 233, p.105759

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. https://doi.org/10.1016/j.fishres.2023.106642

Sfiligoi, I., Bradley, D. C., Holzman, B., Mhashilkar, P., Padhi, S., Wurthwein, F. 2009. The Pilot Way to Grid Resources Using glideinWMS. 2009. WRI World Congress on Computer Science and Information Engineering, Vol. 2, pp. 428–432. doi:10.1109/CSIE.2009.950.

Szuwalski, C.S., Thorson, J.T., 2017. Global fishery dynamics are poorly predicted by classical models. Fish and Fisheries 18, 1085–1095.

Shelton AO, Mangel M. Estimating von Bertalanffy parameters with individual and environmental variations in growth. J Biol Dyn. 2012;6 Suppl 2:3-30. doi: 10.1080/17513758.2012.697195. Epub 2012 Jun 28. PMID: 22882022.

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

Szuwalski, C., 2022. Estimating time-variation in confounded processes in population dynamics modeling: A case study for snow crab in the eastern Bering Sea. Fisheries Research 251, 106298. https://doi.org/10.1016/j.fishres.2022.106298

Taylor, I.G., Methot, R.D., 2013. Hiding or dead? A computationally efficient model of selective fisheries mortality. Fisheries Research 142, 75–85. https://doi.org/10.1016/j.fishres.2012.08.021

Tempel, D.J., Peery, M.Z., Gutiérrez, R.J., 2014. Using Integrated Population Models to Improve Conservation Monitoring: California Spotted Owls as a Case Study. Ecological Modelling 289, 86–95. https://doi.org/10.1016/j.ecolmodel.2014.07.005

Then, A.Y., Hoenig, J.M., Hall, N.G., Hewitt, D.A., Handling editor: Ernesto Jardim, 2015. Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES Journal of Marine Science 72, 82–92. https://doi.org/10.1093/icesjms/fsu136

Thorson, J.T., Munch, S.B., Cope, J.M., Gao, J., 2017. Predicting life history parameters for all fishes worldwide. Ecological Applications 27, 2262–2276. https://doi.org/10.1002/eap.1606

Thorson, J.T., Monnahan, C.C., Hulson, P.-J.F., 2023. Data weighting: An iterative process linking surveys, data synthesis, and population models to evaluate mis-specification. Fisheries Research 266, 106762. https://doi.org/10.1016/j.fishres.2023.106762

Trijoulet, V., Albertsen, C.M., Kristensen, K., Legault, C.M., Miller, T.J., Nielsen, A., 2023. Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fisheries Research 257, 106487. https://doi.org/10.1016/j.fishres.2022.106487

Von Bertalanffy, L. (1957) Quantitative Laws in Metabolism and Growth. Quarterly Review of Biology, 3, 218.

Wang, S.P., Maunder, M.N., 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fisheries Research 192, 41–51. https://doi.org/10.1016/j.fishres.2016.12.005

Winker, H., Carvalho, F., Kapur, M., 2018. JABBA: Just Another Bayesian Biomass Assessment. Fisheries Research 204, 275–288. https://doi.org/10.1016/j.fishres.2018.03.010

Winker, H., Carvalho, F., Thorson, J.T., Kell, L.T., Parker, D., Kapur, M., Sharma, R., Booth, A.J., Kerwath, S.E., 2020. JABBA-Select: Incorporating life history and fisheries' selectivity into surplus production models. Fisheries Research. https://doi.org/10.1016/j.fishres.2019.105355

Zhang, D., O'Conner, N.E., Simpson, A.J., Cao, C., Little, S., Wu, B., 2022. Coastal fisheries resource monitoring through A deep learning-based underwater video analysis. Estuarine, Coastal and Shelf Science 269, 107815. https://doi.org/10.1016/j.ecss.2022.107815

7. Tables

Table 1. Key systems and observation processes and parameter values.

Parameter	**Value**
Natural mortality, M (yr^{-1})	0.2
Reference age, A_{min} (yr)	0
Maximum age, A_{max} (yr)	25
Length at A_{min}, L_{Amin} (cm)	20.5
Length at A_{max}, L_{Amax} (cm)	135.3
Growth rate, k (yr^{-1})	0.19
CV of length < L_{Amin}	0.10
CV of length < L_{Amax}	0.08
Length-weight coefficient	6.8e-6
Length-weight exponent	3.101
Length at 50% maturity, L_{mat50} (cm)	38.18
Slope of maturity ogive	-0.276
Unfished recruitment (Log R_0)	19.0
Spawner-recruitment steepness (h)	0.65
Catchability (Log q)	0.045
Length selectivity for fishery*	50.8, -3, 5.08, 6.99, -999, 999
Length selectivity for survey*	41.8, -4, 4.97, 6.49, -99, 99

*Values for parameterization of double-normal selectivity curve; see Figure S2. For details, see Methot and Wetzel (2013).

Table 2. Percentage of runs (values) that passed each diagnostic across the F *(row groups) and recruitment scenarios (column groups) for various numbers of misspecifications (columns). The values in the cells and the cell color denote the percentage of replicates that passed the tests, ranging from 0% (dark red) to 100% (dark green). Not all diagnostics have published or widely used quantitative thresholds, so pass rates were not obtainable for every diagnostic examined.*

	sigma R = 0.1					**sigma R = 0.4**					**sigma R = 1.0**					
Diagnostic	**0**	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**	
% Converged	100	94	96	98	100	100	97	96	97	100	100	98	97	98	100	**Low F**
Mohn's Rho (SSB)	100	97	97	95	94	94	95	95	97	94	75	76	82	86	81	
Mohn's Rho (F)	94	97	95	94	94	94	95	97	95	100	75	76	81	86	81	
Hindcast Fishery Age	56	55	54	51	50	69	68	66	63	62	75	73	74	78	81	
Hindcast Survey Age	56	53	45	30	12	62	61	51	40	38	62	59	58	54	50	
Hindcast Survey Bio	75	58	57	51	50	50	58	49	52	44	75	73	68	70	75	
Hindcast Fishery Len	88	77	76	78	81	62	68	72	74	75	62	65	66	67	69	
Hindcast Survey Len	62	53	42	38	38	62	56	51	50	50	75	70	66	62	62	
Runs Test Fishery Age	100	93	95	95	94	94	95	93	94	94	94	92	89	89	81	
Runs Test Fishery Len	100	100	99	98	100	100	98	97	100	100	100	98	96	94	94	
Runs Test Survey Age	81	80	79	78	69	100	100	96	89	75	100	97	90	87	81	
Runs Test Survey Bio	100	98	93	89	94	100	100	100	100	94	94	94	94	94	94	
Runs Test Survey Len	94	95	96	95	94	100	95	95	95	94	100	95	91	89	75	
% Converged	100	77	68	55	56	100	81	66	64	69	100	64	65	53	50	**High F**
Mohn's Rho (SSB)	100	98	98	97	100	100	100	100	100	100	100	98	98	100	100	
Mohn's Rho (F)	100	98	98	97	100	100	100	100	100	100	94	95	98	100	100	
Hindcast Fishery Age	62	59	63	51	44	62	62	65	61	64	81	78	74	76	88	
Hindcast Survey Age	56	55	42	37	22	56	56	41	41	27	62	56	47	47	50	
Hindcast Survey Bio	69	67	68	74	67	62	60	63	51	45	50	56	55	50	38	
Hindcast Fishery Len	69	65	60	63	67	56	60	57	54	64	69	66	60	62	62	
Hindcast Survey Len	50	47	45	43	44	69	62	65	61	55	75	73	69	50	25	
Runs Test Fishery Age	100	98	94	94	100	94	98	98	100	91	88	93	95	97	100	
Runs Test Fishery Len	94	94	95	91	89	100	98	98	100	100	100	100	98	97	100	
Runs Test Survey Age	94	92	85	86	89	100	92	81	73	55	100	90	84	76	62	
Runs Test Survey Bio	100	98	100	100	100	100	100	100	100	100	100	100	98	97	100	
Runs Test Survey Len	94	96	94	91	78	100	96	94	95	100	100	93	90	79	62	

Table 3. Tentative good practices and precautions for applying diagnostic tests to assessments. ***This table is meant to be taken as an ordered guide; should a model fail the diagnostic "good practice" for a given row, we suggest exploration of alternative model structure(s).***

Diagnostic	**Good Practice**	**Precaution**
Plausibility	Contextualize model in ecological, life-history and fishery dynamics (realism)	Risk of model over-complication (e.g., too many or improper time-varying processes, Szuwalski 2017; Fisch 2023,) Particularly difficult when working with multi-species and multi-area assessments
Convergence and Check for Global Solution	Final gradient below pre-specified minimum (e.g., 1E-4); Hessian matrix is invertible; No parameters are on bound Consider Bayesian approaches when applicable	Avoid massaging data or exhaustive chain lengths (jitter, MCMC) to force convergence; set terms of analyses ahead-of-time
Goodness of fit		
Residual Diagnostics	Visual inspection of residuals and model fits	The p-value of 0.05 (runs test) might be too low; RMSE cutoff of 30% might be too high Tests appear more sensitive when applied to compositional data than indices Beware small time series Consider One-Step-Ahead residuals for compositional data (Trijoulet et. al. 2022)
Model Consistency		
R0 likelihood profile	Profile over key model parameters (R_0, M and steepness if applicable); Check for minima outside of 95% CI of base model; Evaluate data conflicts and likelihood surface	Consider how prior likelihoods are included; Model specification and data weighting can impact behavior (Wang et al., 2017)
Age structured production model	Explore when there are multiple data sources, especially for compositions	Recruitment-driven models (e.g. short-lived species, low recruitment variability, and/or low exploitation history) might have poorly defined production functions; Biomass scale might be poorly informed when fishing mortality is low (Minte-Vera et al. 2022)
Retrospective analysis	Visually inspect retrospective	Rho within fixed threshold cannot rule

	patterns;
	Rose approach, resource permitting (Legault, 2020);
	Consider post-sample rho (Breivik et al. 2023);
	Consider model-specific confidence intervals for rho (Miller and Legault, 2017
		out parameter misspecification;
		Consider time-varying processes and data weighting;
		Reference points can remain biased even when retrospective patterns disappear (Szuwalski et al. 2018)

Prediction Skill	Consider leave-one-out cross validation, especially when time series are sparse or few	MASE criterion within threshold cannot rule out parameter misspecification; more research needed
Recruitment trend	Visually inspect recruitment deviates/calculate quantitative metrics for trend and non-randomness in the deviates (Merino et al., 2022)	Models with significant linear trend in the recruitment deviates are likely to be mis-specified; the *absence* of linear trend is not evidence that the model is correctly specified.

8. Figures

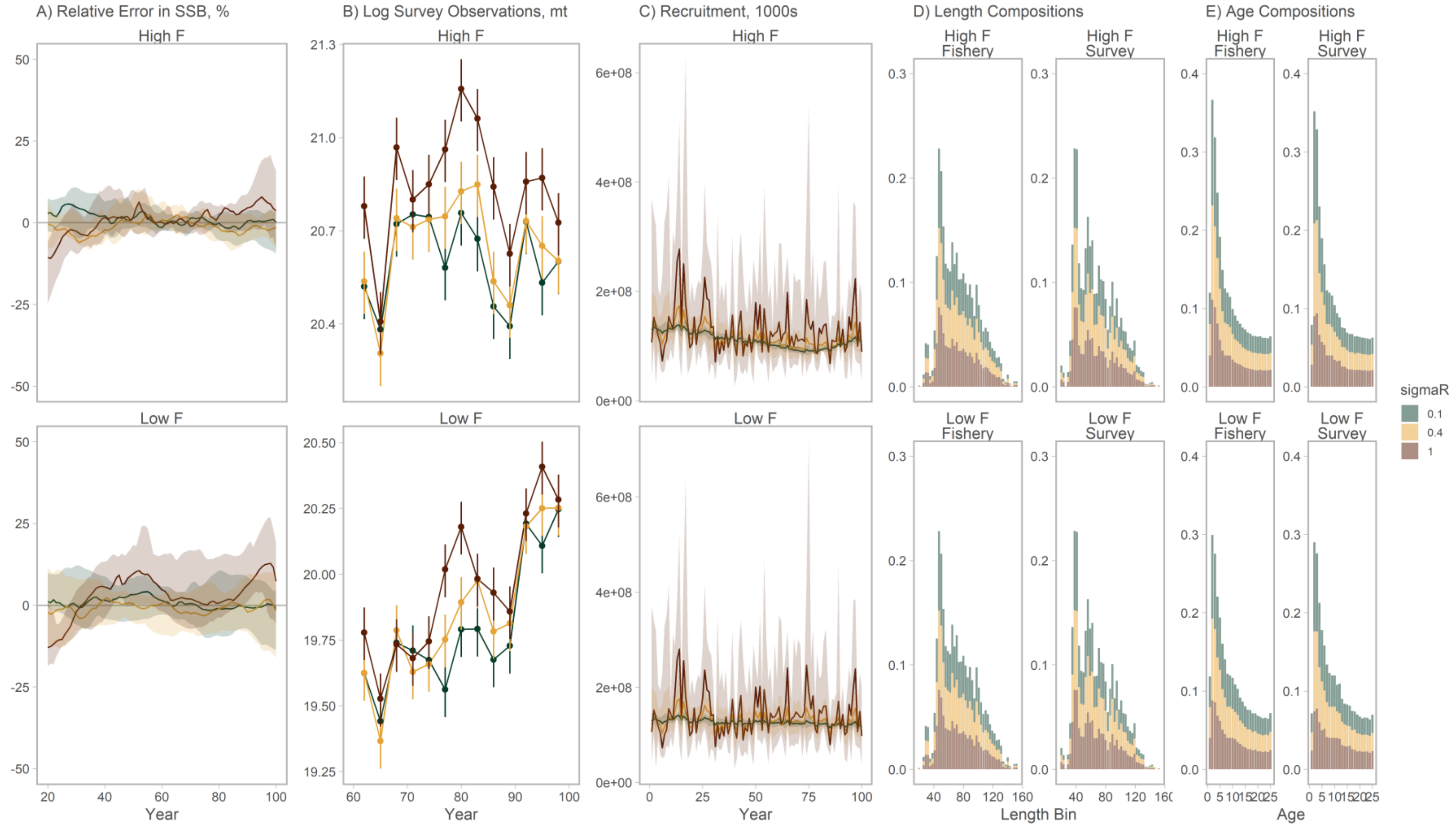


Figure 1. A) Relative error (%) in depletion for estimation models with no misspecifications. B) Survey observations from the operating model (points and lines) with 95% simulation intervals (values are summarized across all OM replicates). C) Annual recruitment in 1000s of individuals. D) Observed length compositions, aggregated across time, for the fishery and survey fleets. E) Observed age compositions, aggregated across time, for the fishery and survey fleets. In all plots, colors correspond to the recruitment variability scenarios. The shaded ribbons in A) and C) correspond to the 95% simulation interval

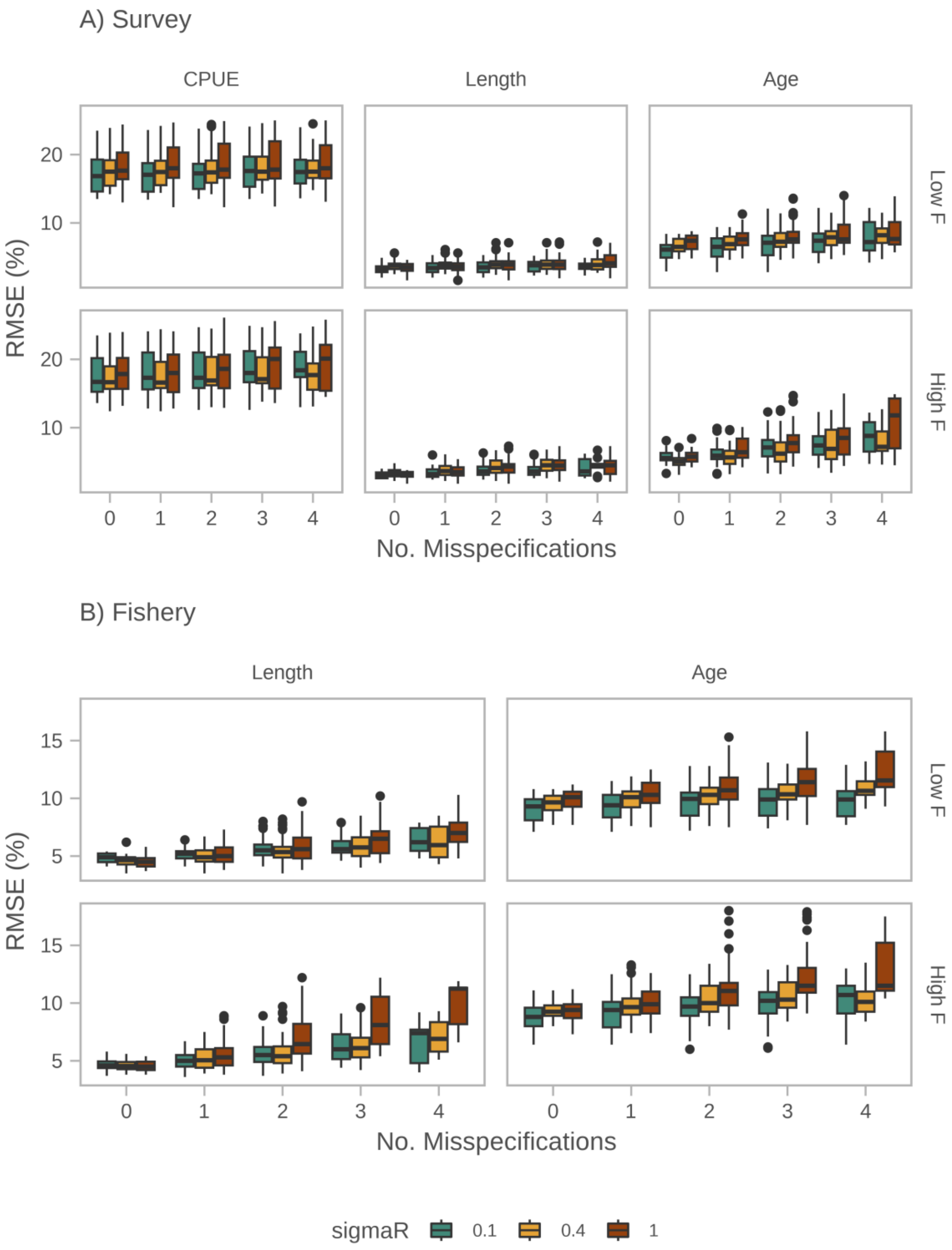


Figure 2. Boxplots of RMSE of A) survey indices of abundance, length and age composition data and B) fishery length and age composition data (bottom) for two levels of fishing mortality (rows). The x-axis represents the number of misspecifications present in the estimation method (0 mis-specifications corresponds to the correct estimator). Colors correspond to the value of recruitment variability used in the OM.

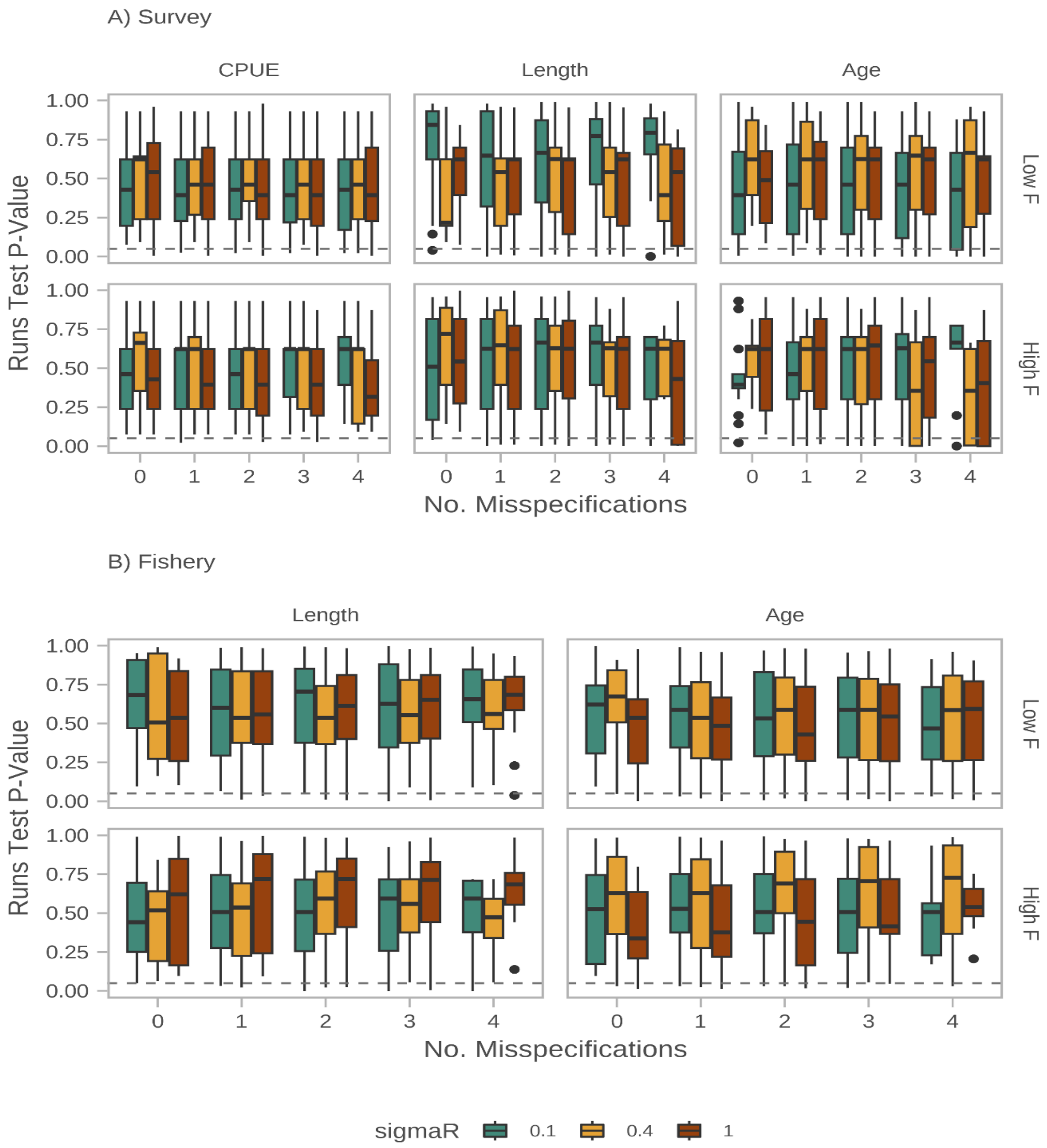


Figure 3. Box plots of p-values from Runs test for A) survey indices of abundance, length and age composition data and B) fishery length and age composition data (bottom) for two levels of fishing mortality (rows). The traditional interpretation of this test is that p-values greater than 0.05 (dashed line) indicate no evidence to reject the null hypothesis that residuals are normally distributed (thus values above the line "pass" the test). The x-axis represents the number of misspecifications present in the estimation method (0 mis-specifications corresponds to the correct estimator). Colors correspond to the value of recruitment variability used in the OM.

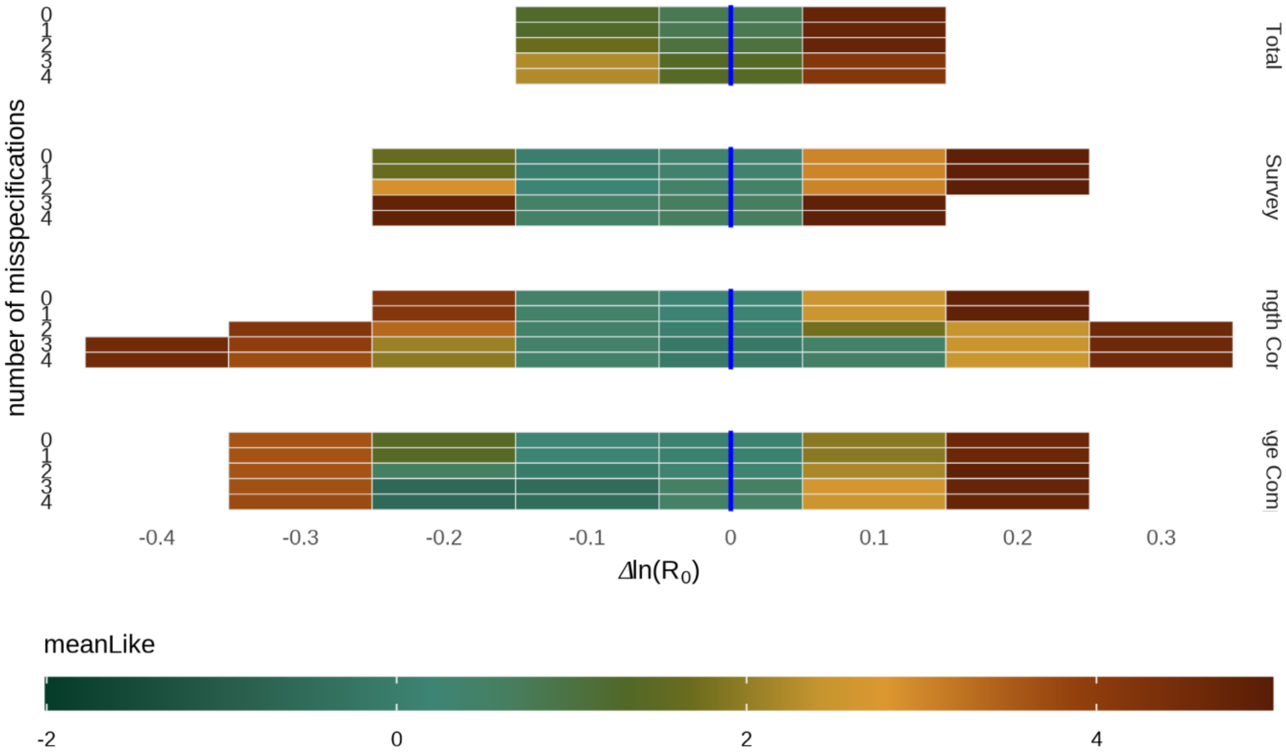


Figure 4. Likelihood profiles for log(R0) shown for a subset of estimation methods with zero through four misspecifications. Each panel corresponds to either the total likelihood (top), or survey, length or age composition components (bottom three panels). The x-axis has been re-centered to the corresponding MLE from the correct (not mis-specified parameters) estimation method (vertical blue line); profiles have been filtered to only display model runs with changes in the scaled negative log-likelihood less than 10 units. Green tiles indicate models closer to the minimum negative log-likelihood; red values are higher.

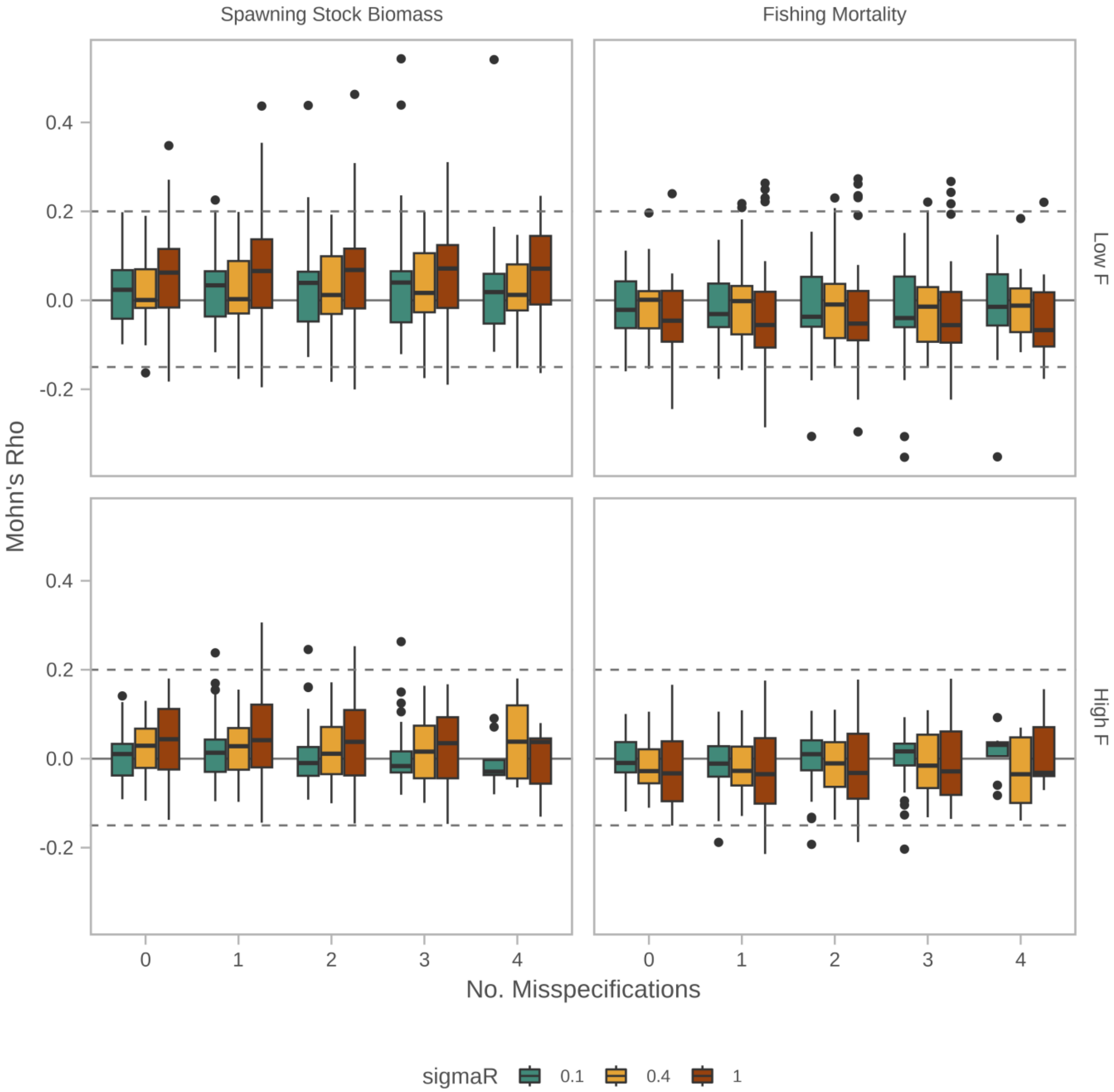


Figure 5. Boxplots of rho from 5-year retrospectives in SSB (left) and fishing mortality (right) for two levels of fishing mortality (rows). The x-axis represents the number of mis-specified parameters in the estimation method (0 misspecifications corresponds to the correct estimator). Colors correspond to the value of recruitment variability used in the OM. Dashed lines indicate the thresholds suggested by Hurtado-Ferro et al. (2015) for poor rho values.

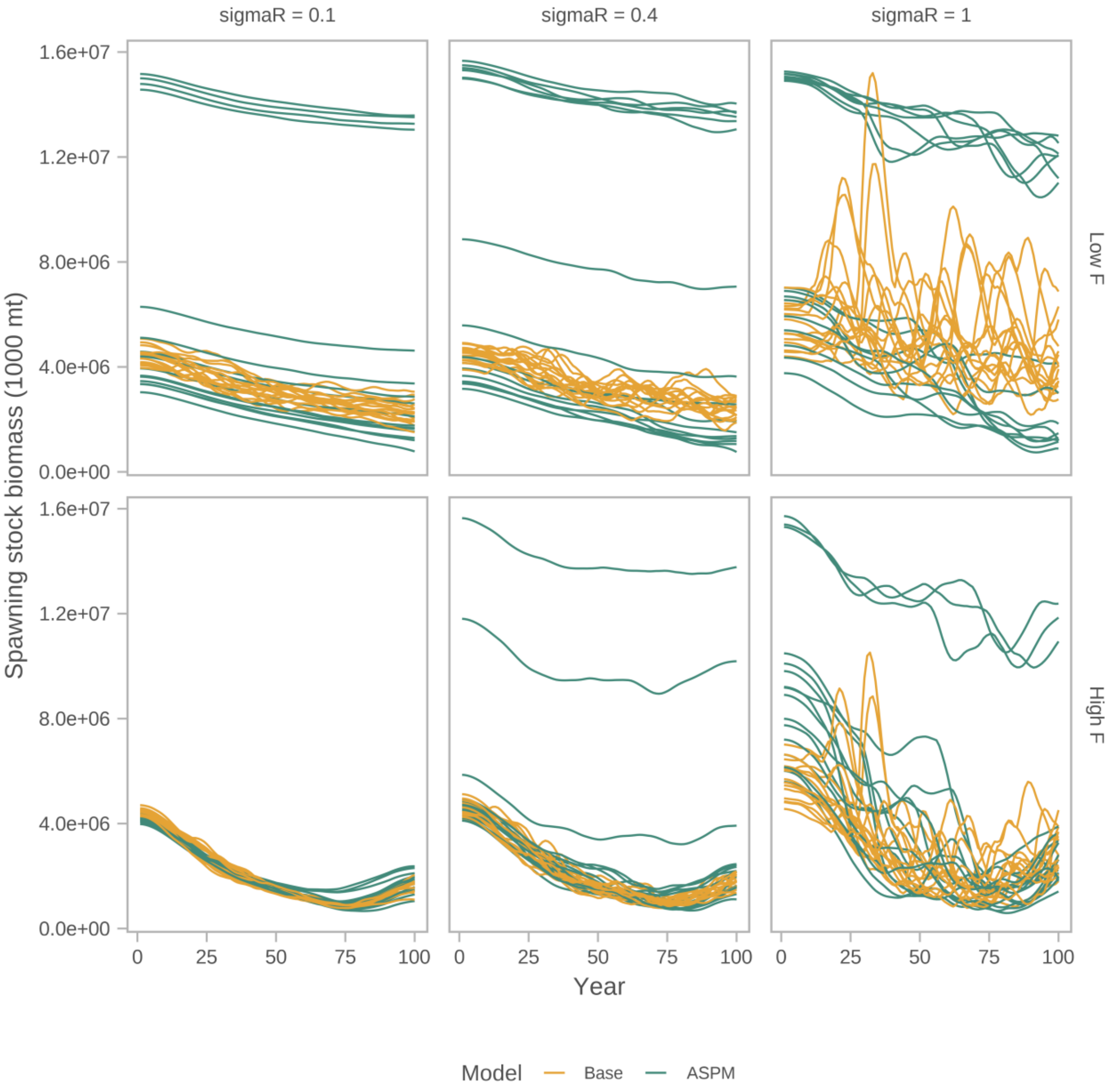


Figure 6. Estimated spawning stock biomass trajectories for correctly specified ASPM EMs (green lines) and correctly specified age-structured integrated model (yellow lines).

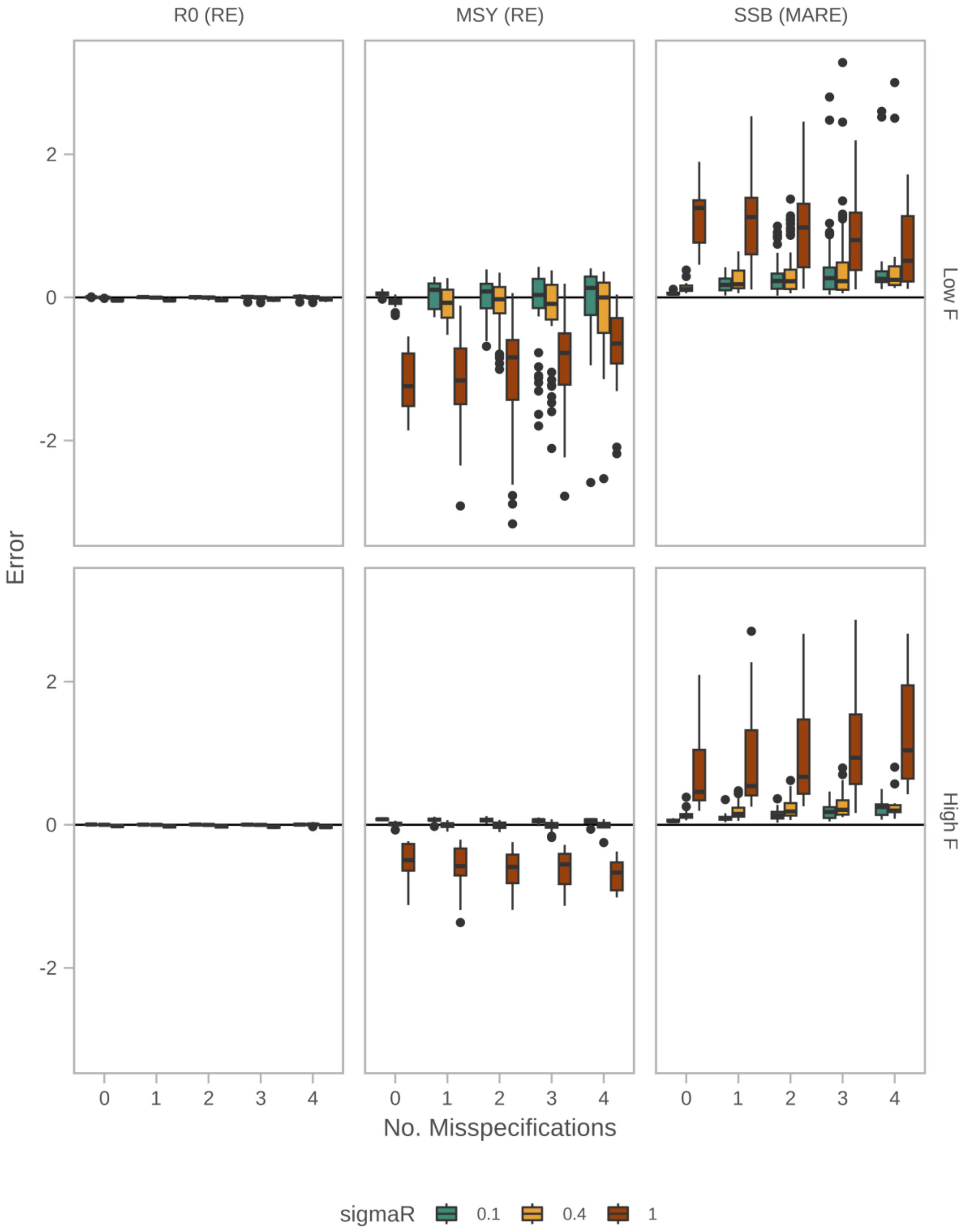


Figure 7. Boxplots of the relative difference in a deterministic recruitment model and full model of R_0, MSY, and MARE of SSB *for two levels of* fishing mortality *(rows). The x-axis represents the number of mis-specifications present in the estimation method (0 misspecifications corresponds to the correct estimator). Colors correspond to the value of recruitment variability used in the OM.*

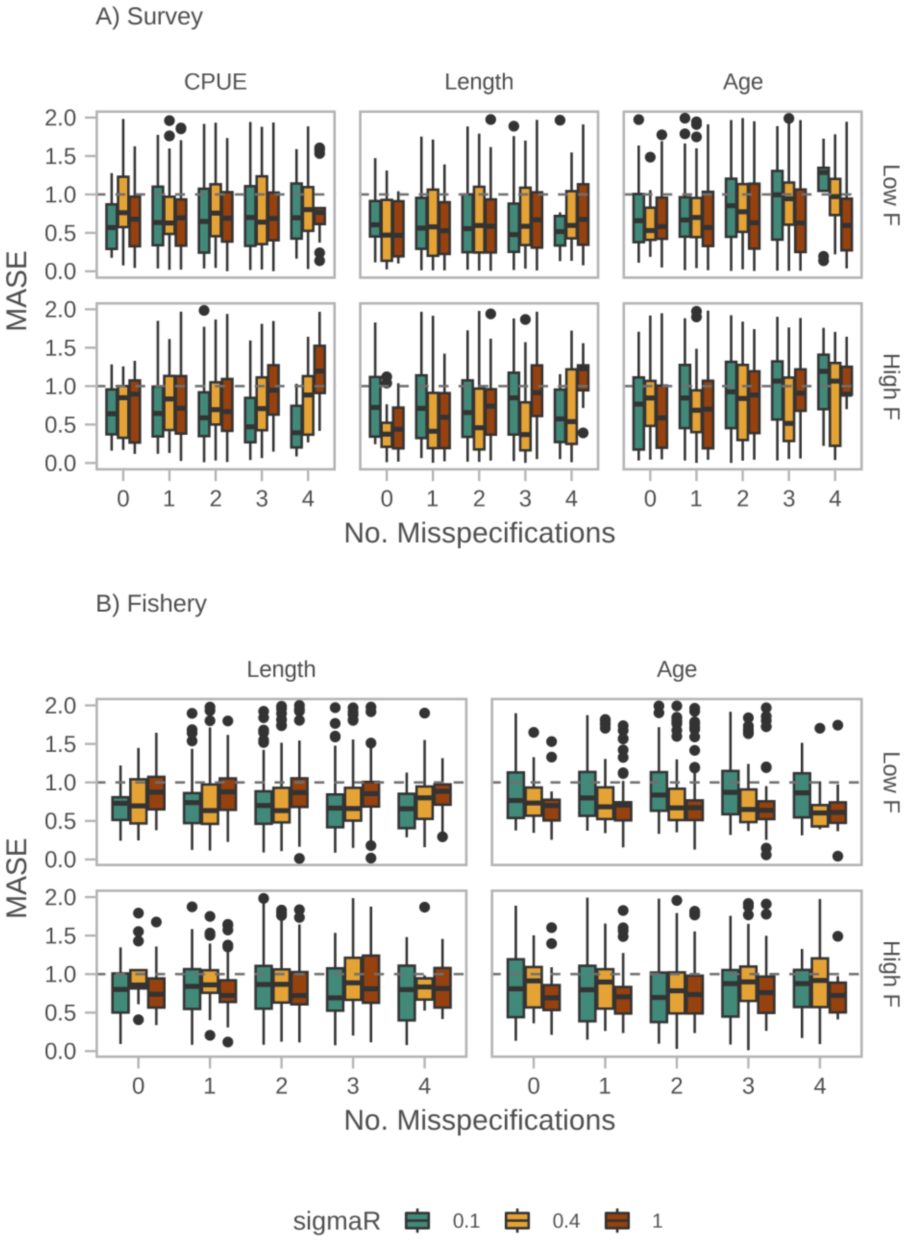


Figure 8. Boxplots of hindcast cross-validation MASE for A) survey indices of abundance, length and age composition data and B) fishery length and age composition data (bottom) for two levels of fishing mortality (rows). The x-axis represents the number of mis-specifications present in the estimation method (0 misspecifications corresponds to the correct estimator). Colors correspond to the value of recruitment variability used in the OM. MASE scores below 1 (dashed line) have greater predictive power than a null model.

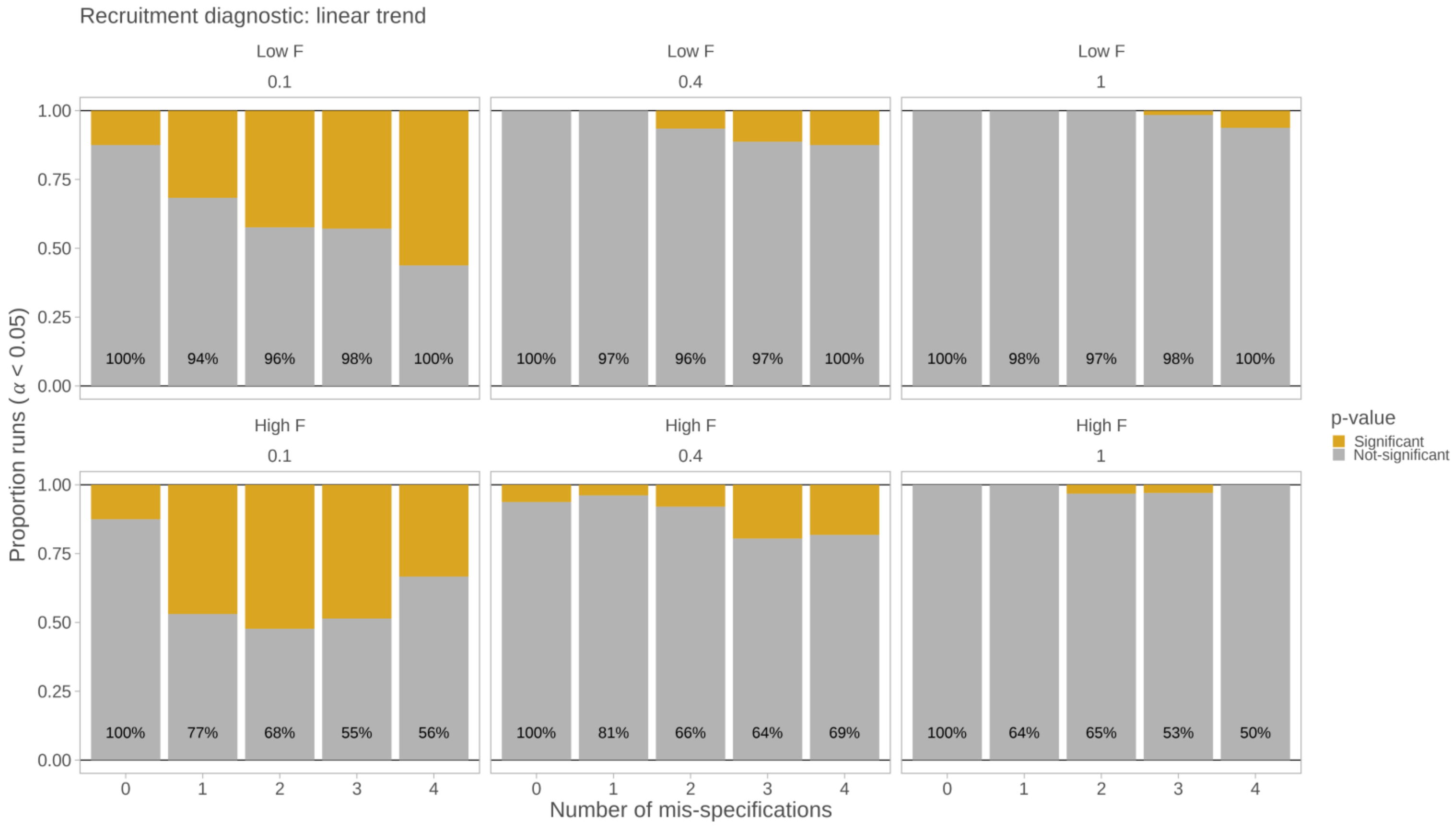


Figure 9. Proportion of converged models with significant results for tests of linear trend in the recruitment deviates by number of mis-specifications, levels of fishing mortality, and assumed recruitment variability. The percentage shown at the bottom of each bar is the percentage of converged models relative to the total number of models run.

9. Supplementary material

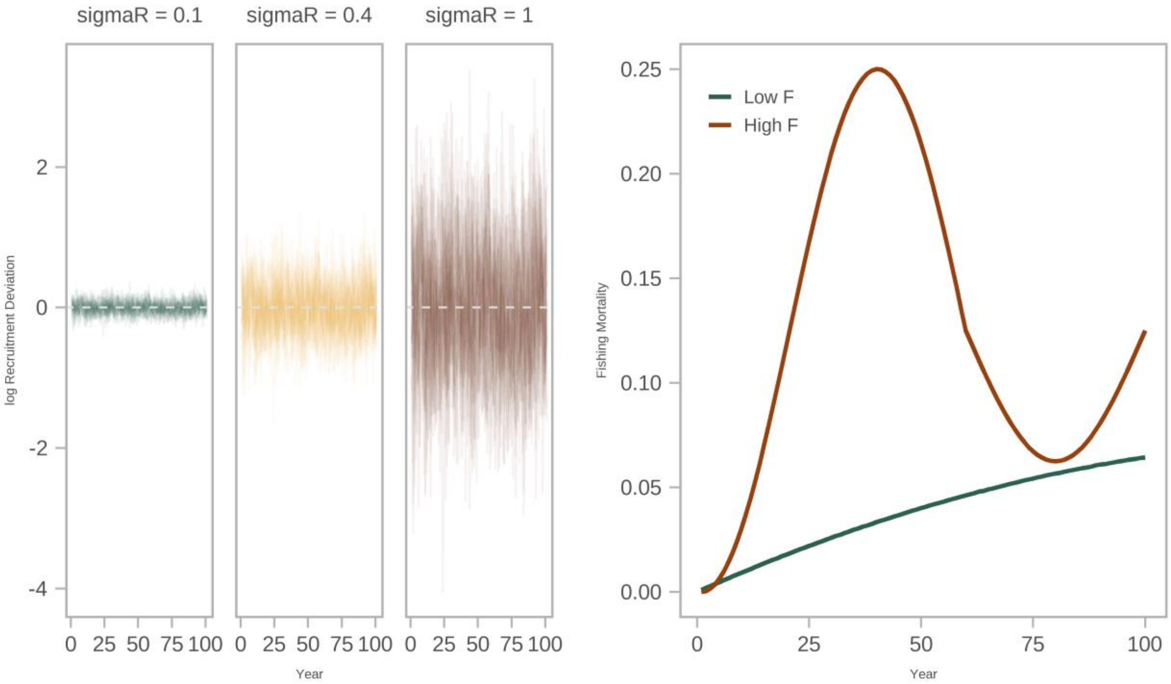


Figure S1. (Left) time series of simulated recruitment deviations at three different levels of variation. (Right) Input fishing mortality vectors for the “low” (green) and “high” fishing mortality scenario.

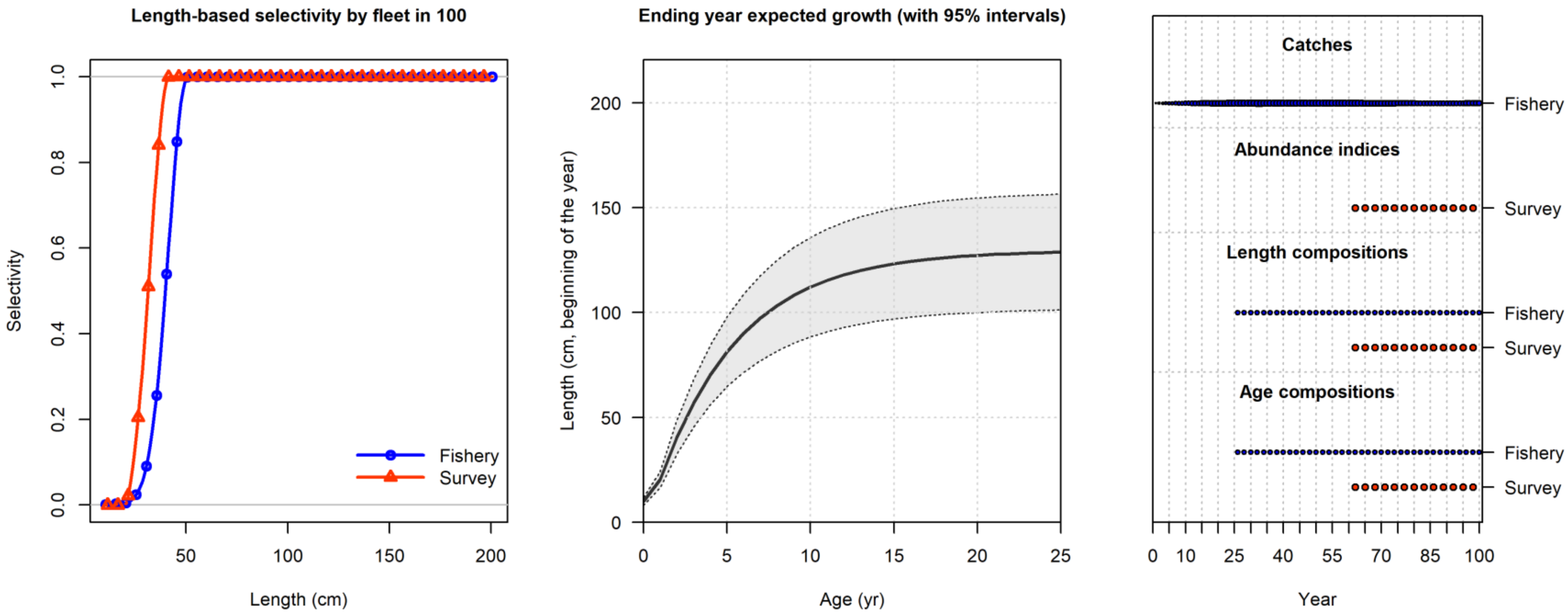


Figure S2. Operating model values for selectivity (left) and growth (center) and data availability by year for the fishery fleet and survey (right). All OM replicates and associated EMs have the same years of data available.

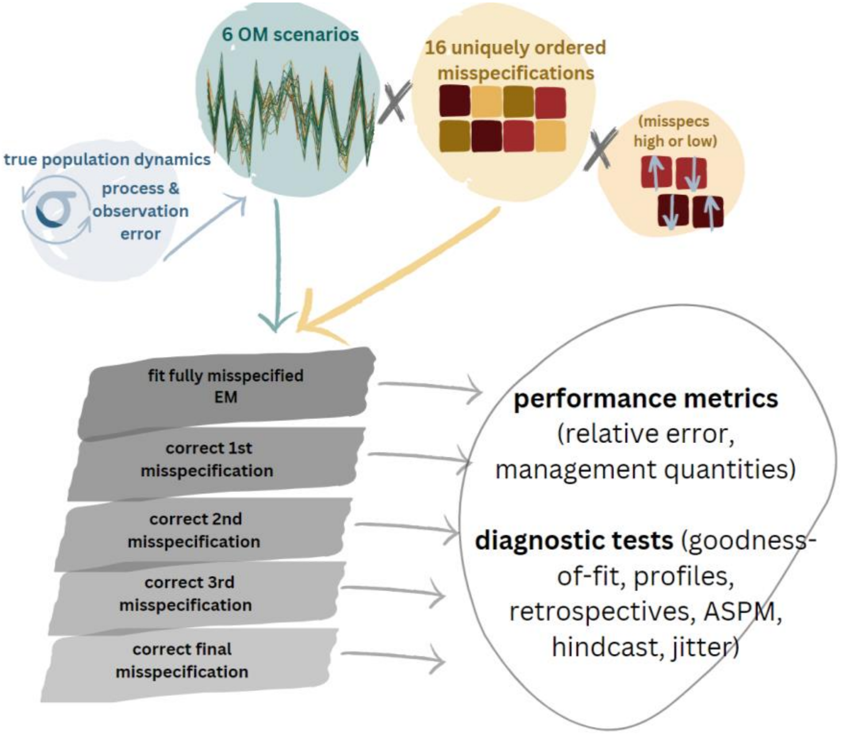

6 OM scenarios
16 uniquely ordered misspecifications
(misspecs high or low)
true population dynamics
process & observation error
fit fully misspecified EM
correct 1st misspecification
correct 2nd misspecification
correct 3rd misspecification
correct final misspecification
performance metrics (relative error, management quantities)
diagnostic tests (goodness-of-fit, profiles, retrospectives, ASPM, hindcast, jitter)

Figure S3. Schematic of the main experimental design. Operating model replicates are made by simulating datasets given process and observation error (catches are assumed known). All uniquely ordered sets of misspecifications (including parameter identity and direction of misspecification) are fit to each replicate and corrected sequentially. For all unique estimation methods, we run a suite of diagnostic tests as described in Carvalho et al. (2021).

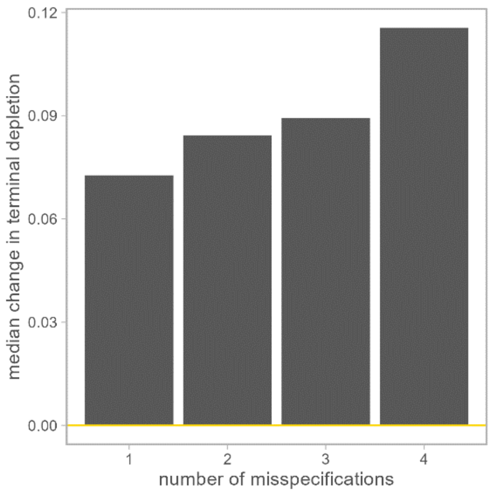


Figure S4. Median change in terminal depletion versus number of misspecifications, for converged results only. This illustrates that the error in terminal depletion values cumulatively increases with increasing misspecifications.

Table S1. Summary of mis-specified values for natural mortality, steepness, selectivity, and asymptotic length identified during the thresholding experiments for unique combinations of fishing mortality and variation in recruitment deviations. The true values are shown in OM Value, and the mis-specified values above and below the true value are shown in the last two columns.

OM Scenario	Parameter	OM Values [bounds]	Value that results in ~10% decrease in terminal SSB	Value that results in ~10% increase in terminal SSB
F=low sigmaR=0.1	Natural Mortality	0.2 [0.01,1.8]	0.188	0.212
F=low sigmaR=0.4			0.188	0.212
F=low sigmaR=1			0.188	0.214
F=high sigmaR=0.1			0.188	0.216
F=high sigmaR=0.4			0.19	0.214
F=high sigmaR=1			0.186	0.214
F=low sigmaR=0.1	Steepness	0.65 [0.2,1]	0.4615	0.988
F=low sigmaR=0.4			0.4615	0.988
F=low sigmaR=1			0.4615	0.988
F=high sigmaR=0.1			0.5915	0.715
F=high sigmaR=0.4			0.585	0.7215
F=high sigmaR=1			0.5915	0.7345
F=low sigmaR=0.1	Length at 50% selectivity	50.8 [5.08,101.6]	44.196	58.42
F=low sigmaR=0.4			43.688	58.42
F=low sigmaR=1			42.672	58.928
F=high sigmaR=0.1			43.688	59.436
F=high sigmaR=0.4			44.196	57.912
F=high sigmaR=1			40.64	62.484
F=low sigmaR=0.1	Asymptotic length	129.5 [6.6,660]	126.72	138.6
F=low sigmaR=0.4			126.72	138.6
F=low sigmaR=1			126.72	138.6
F=high sigmaR=0.1			125.4	139.92
F=high sigmaR=0.4			120.12	138.6
F=high sigmaR=1			122.76	141.24

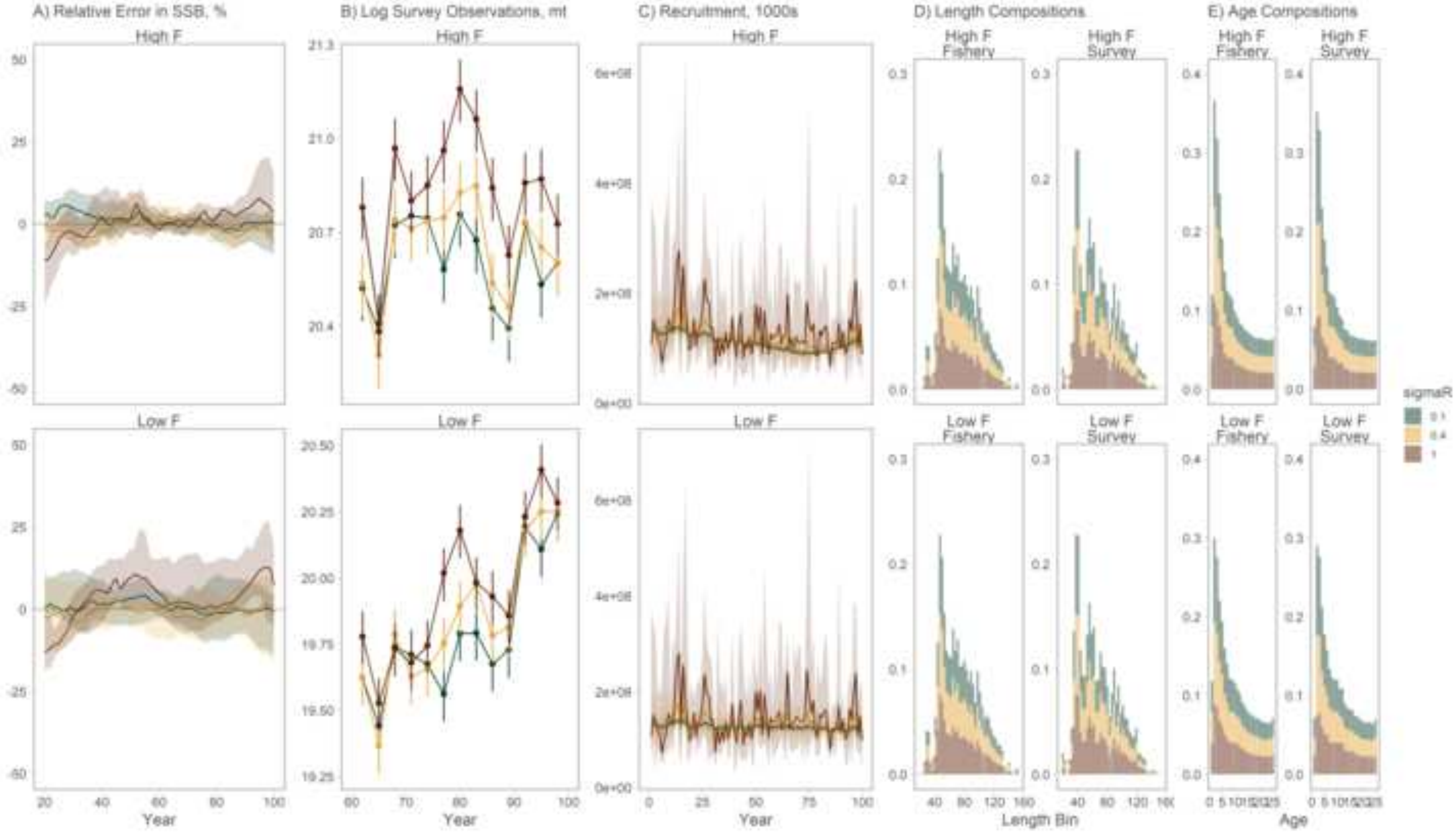
Table 2

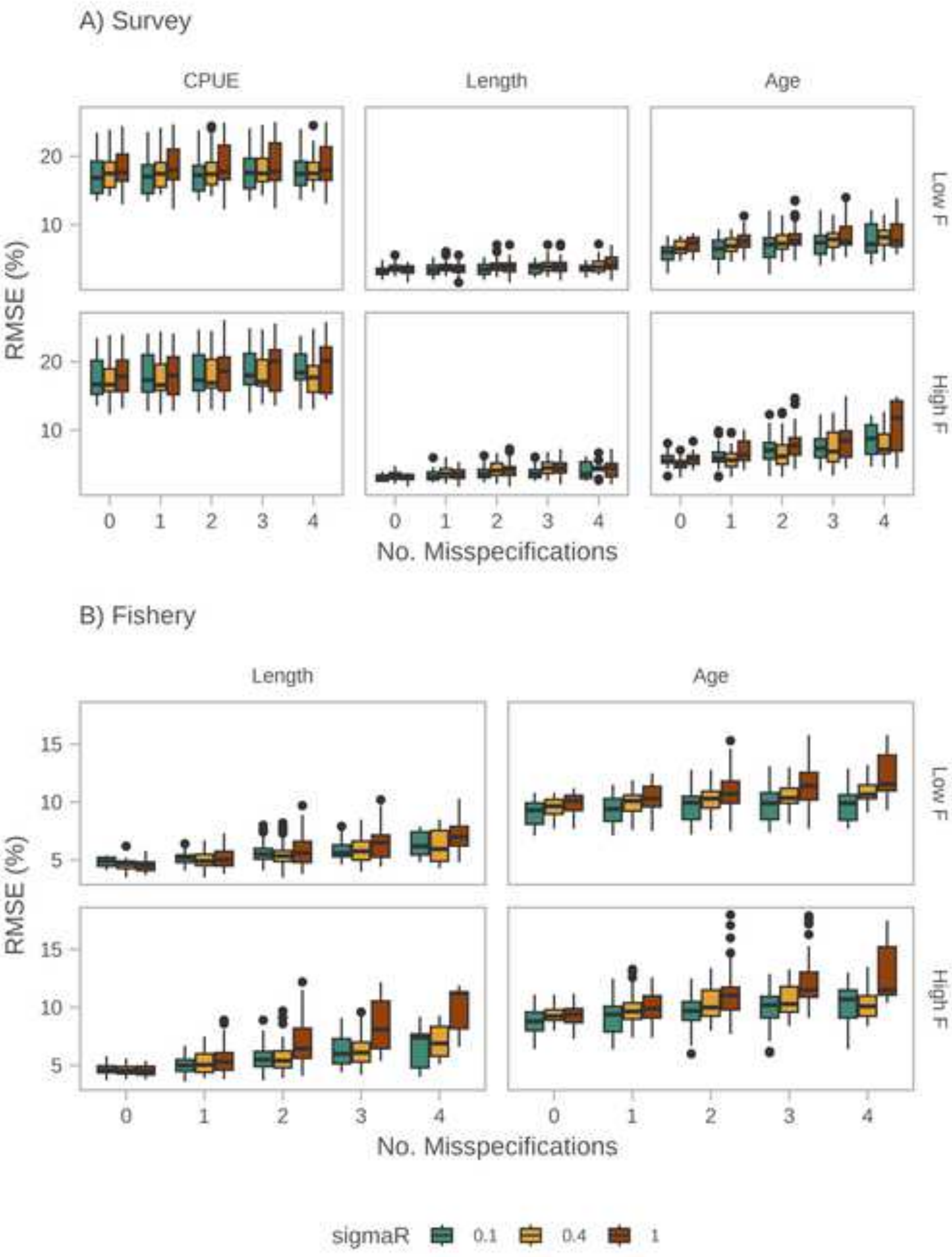
Click here to access/download;Table;table2_reformat.xlsx

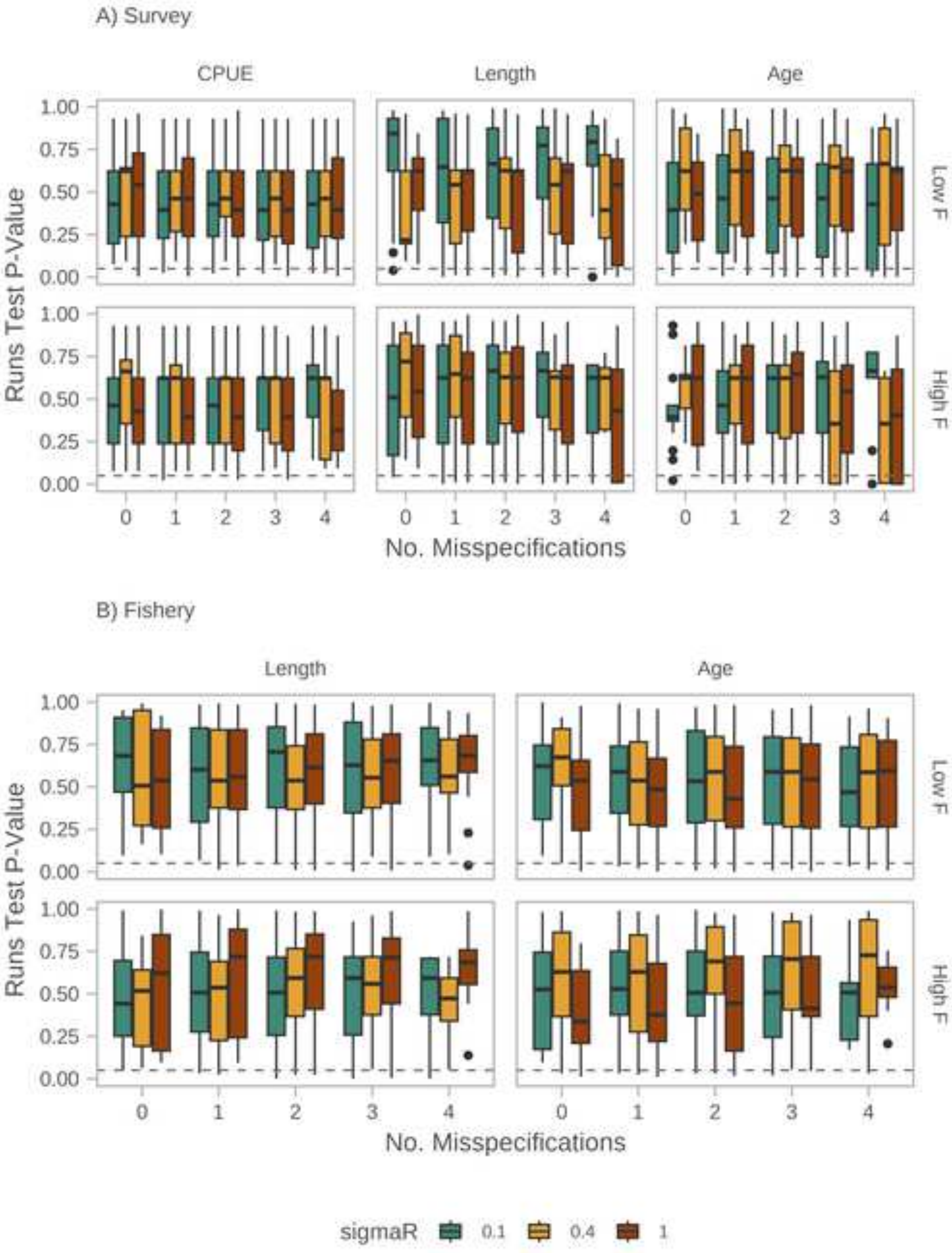
	sigma R = 0.1					sigma R = 0.4					sigma R =		
Diagnostic	0	1	2	3	4	0	1	2	3	4	0	1	2
% Converged	100	94	96	98	100	100	97	96	97	100	100	98	97
Mohn's Rho (SSB)	100	97	97	95	94	94	95	95	97	94	75	76	82
Mohn's Rho (F)	94	97	95	94	94	94	95	97	95	100	75	76	81
Hindcast Fishery Age	56	55	54	51	50	69	68	66	63	62	75	73	74
Hindcast Survey Age	56	53	45	30	12	62	61	51	40	38	62	59	58
Hindcast Survey Bio	75	58	57	51	50	50	58	49	52	44	75	73	68
Hindcast Fishery Len	88	77	76	78	81	62	68	72	74	75	62	65	66
Hindcast Survey Len	62	53	42	38	38	62	56	51	50	50	75	70	66
Runs Test Fishery Age	100	93	95	95	94	94	95	93	94	94	94	92	89
Runs Test Fishery Len	100	100	99	98	100	100	98	97	100	100	100	98	96
Runs Test Survey Age	81	80	79	78	69	100	100	96	89	75	100	97	90
Runs Test Survey Bio	100	98	93	89	94	100	100	100	100	94	94	94	94
Runs Test Survey Len	94	95	96	95	94	100	95	95	95	94	100	95	91
% Converged	100	77	68	55	56	100	81	66	64	69	100	64	65
Mohn's Rho (SSB)	100	98	98	97	100	100	100	100	100	100	100	98	98
Mohn's Rho (F)	100	98	98	97	100	100	100	100	100	100	94	95	98
Hindcast Fishery Age	62	59	63	51	44	62	62	65	61	64	81	78	74
Hindcast Survey Age	56	55	42	37	22	56	56	41	41	27	62	56	47
Hindcast Survey Bio	69	67	68	74	67	62	60	63	51	45	50	56	55
Hindcast Fishery Len	69	65	60	63	67	56	60	57	54	64	69	66	60
Hindcast Survey Len	50	47	45	43	44	69	62	65	61	55	75	73	69
Runs Test Fishery Age	100	98	94	94	100	94	98	98	100	91	88	93	95
Runs Test Fishery Len	94	94	95	91	89	100	98	98	100	100	100	100	98
Runs Test Survey Age	94	92	85	86	89	100	92	81	73	55	100	90	84
Runs Test Survey Bio	100	98	100	100	100	100	100	100	100	100	100	100	98
Runs Test Survey Len	94	96	94	91	78	100	96	94	95	100	100	93	90

1.0		
3	4	
98	100	Low F
86	81	
86	81	
78	81	
54	50	
70	75	
67	69	
62	62	
89	81	
94	94	
87	81	
94	94	
89	75	
53	50	High F
100	100	
100	100	
76	88	
47	50	
50	38	
62	62	
50	25	
97	100	
97	100	
76	62	
97	100	
79	62	

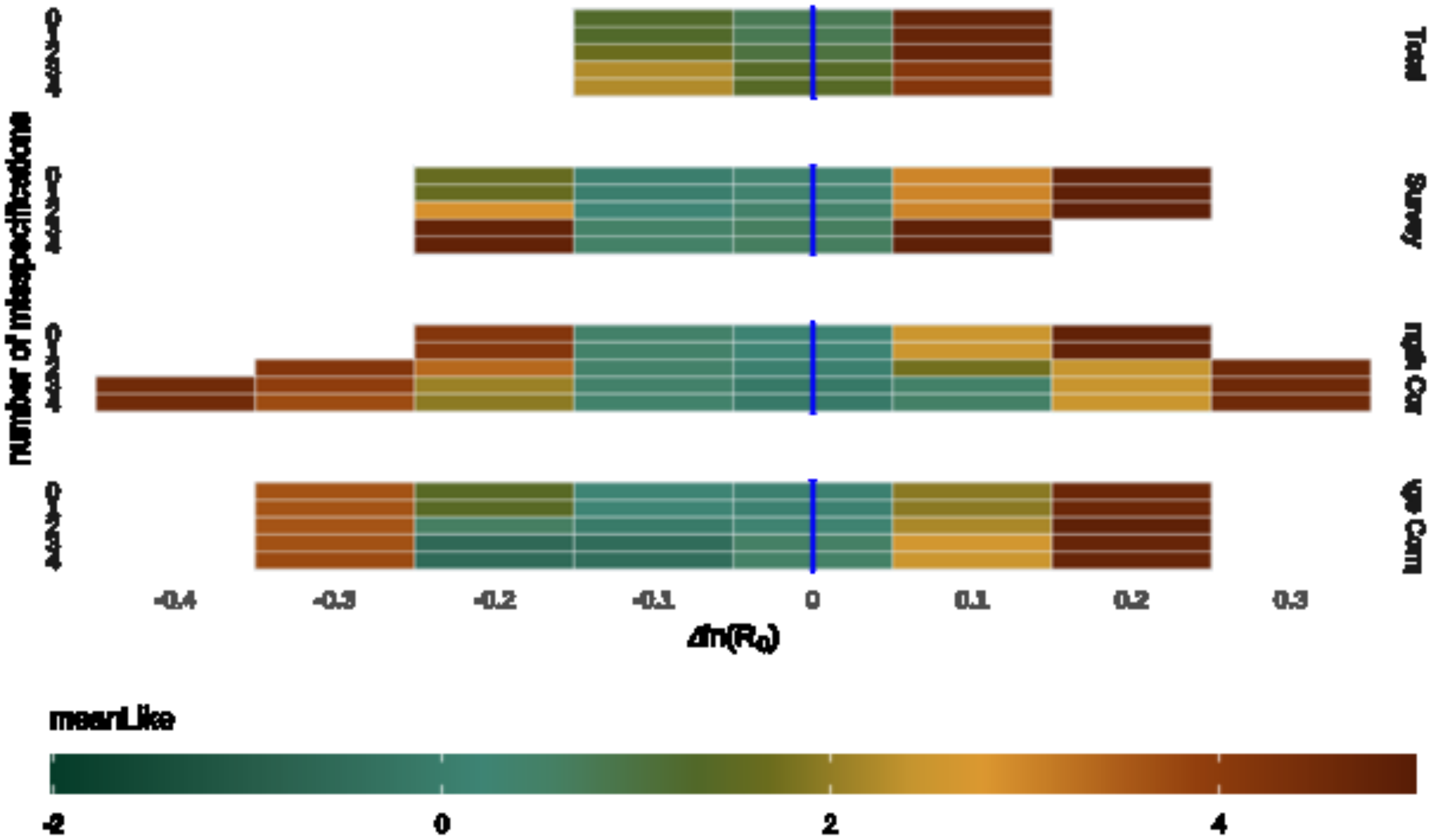
Figure 1

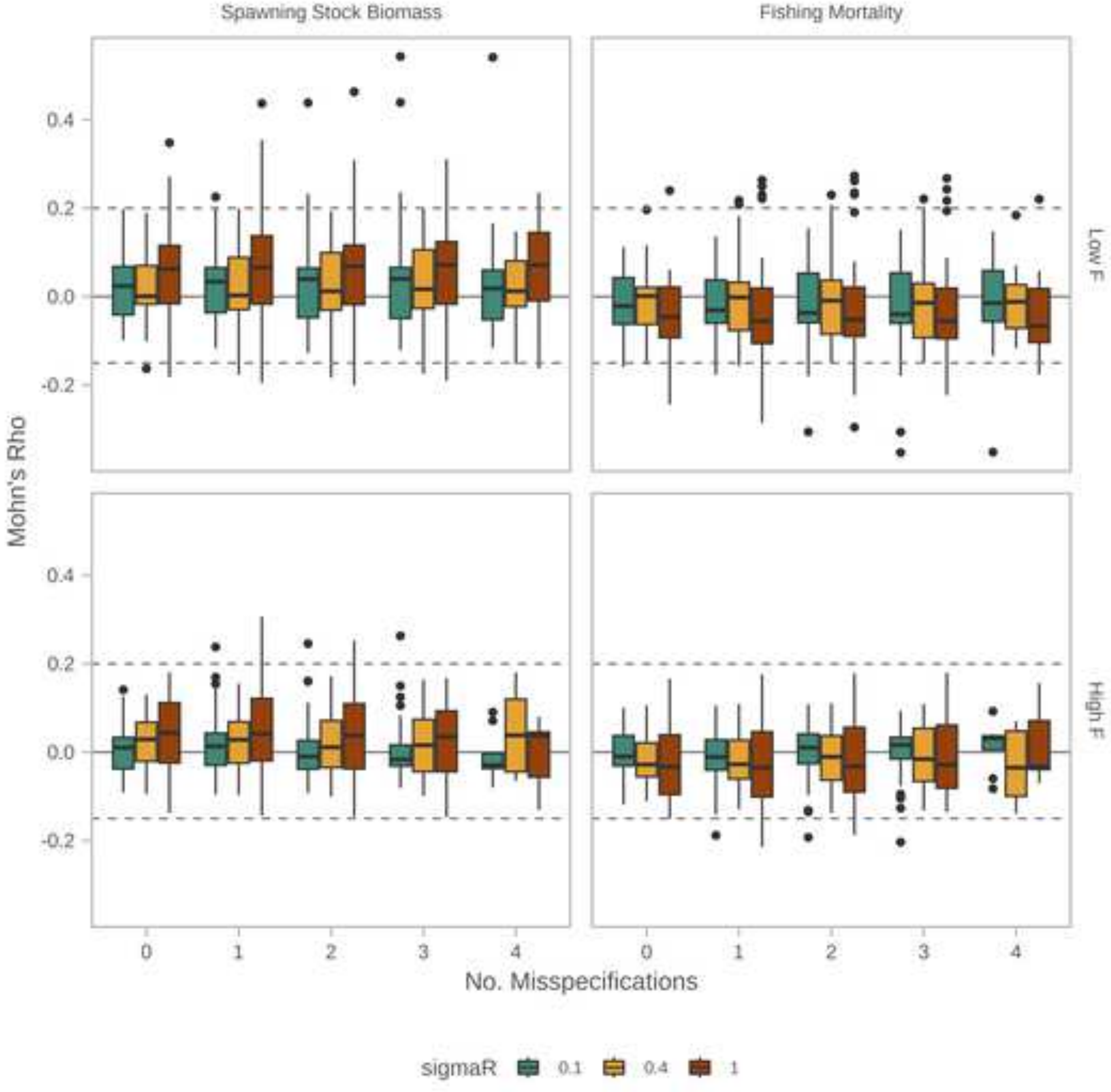

A) Relative Error in SSB, %
B) Log Survey Observations, mt
C) Recruitment, 1000s
D) Length Compositions
E) Age Compositions
High F
Low F
High F Fishery
High F Survey
Low F Fishery
Low F Survey
50
25
0
-25
-50
21.3
21.0
20.7
20.4
20.50
20.25
20.00
19.75
19.50
19.25
6e+08
4e+08
2e+08
0e+00
0.3
0.2
0.1
0.0
0.4
20
40
60
80
100
60
70
80
90
100
0
25
50
75
100
40
80
120
160
0 5 10 15 20 25
Year
Length Bin
Age
sigmaR
0.1
0.4
1

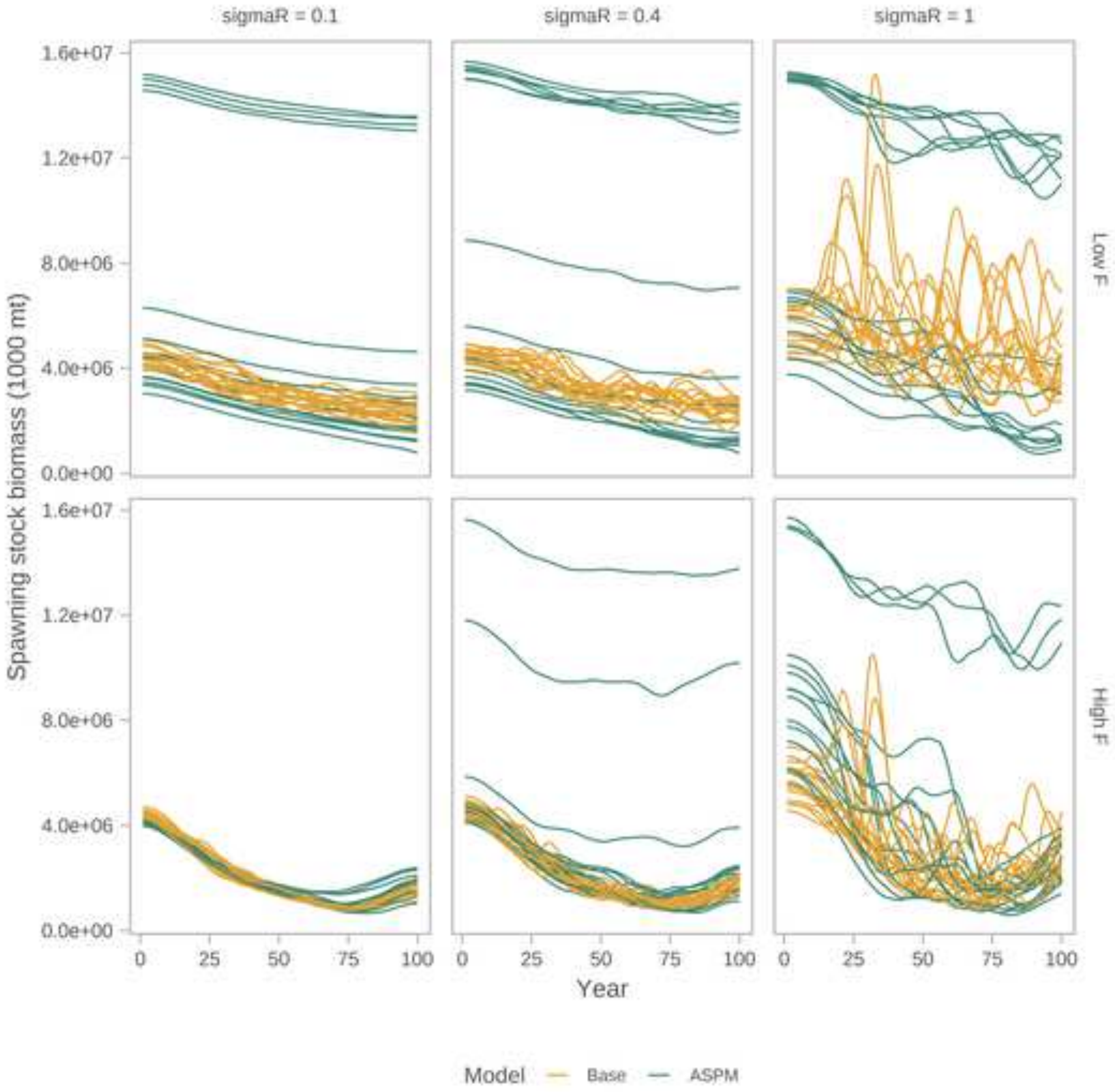

A) Survey
CPUE
Length
Age
Low F
High F
RMSE (%)
20
10
0
1
2
3
4
No. Misspecifications
B) Fishery
Length
Age
15
10
5
sigmaR
0.1
0.4
1

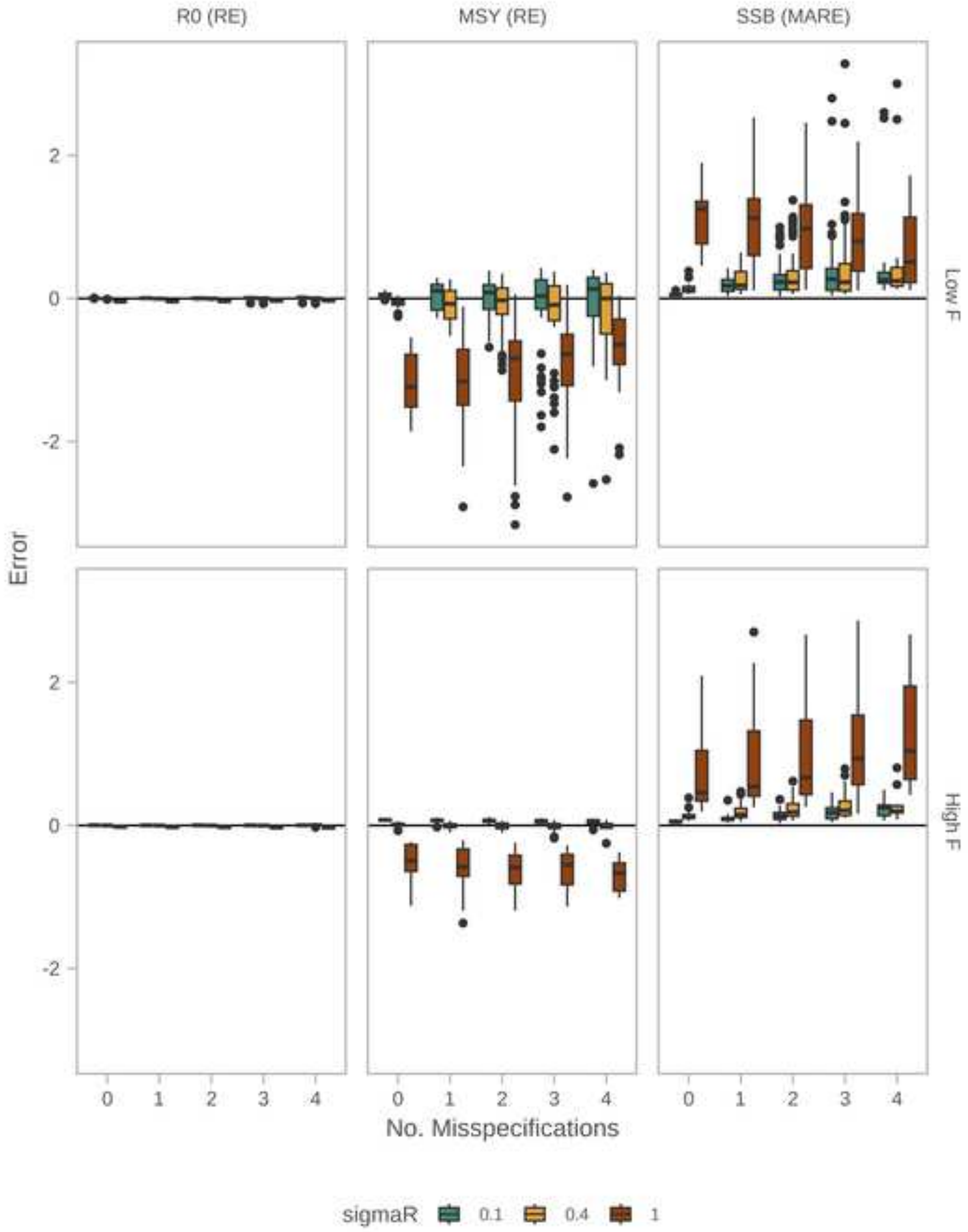

A) Survey
CPUE
Length
Age
Low F
High F
Runs Test P-Value
1.00
0.75
0.50
0.25
0.00
0
1
2
3
4
No. Misspecifications
B) Fishery
Length
Age
Low F
High F
Runs Test P-Value
No. Misspecifications
sigmaR
0.1
0.4
1

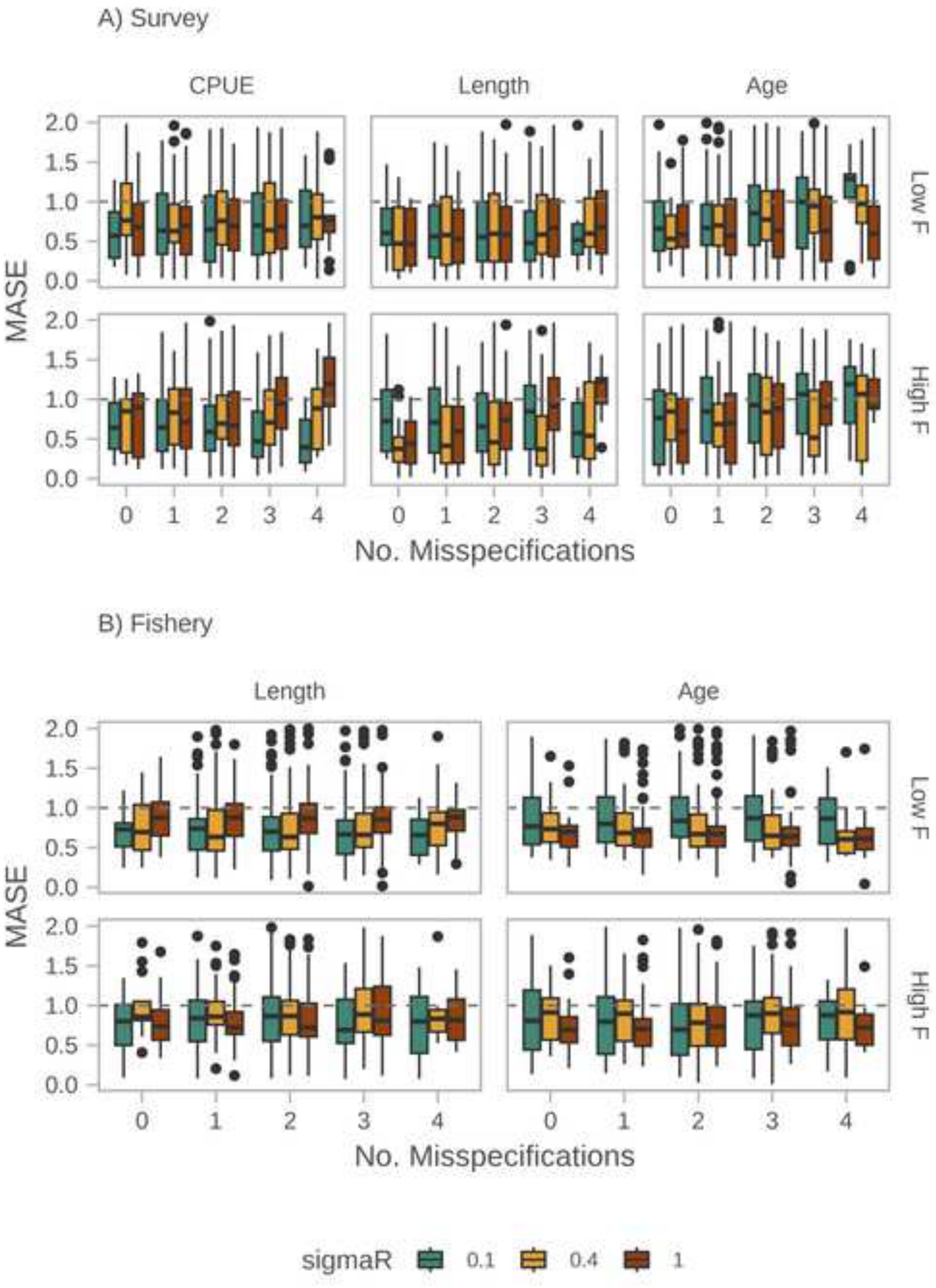
Click here to access/download;Figure;Figure4_update.png

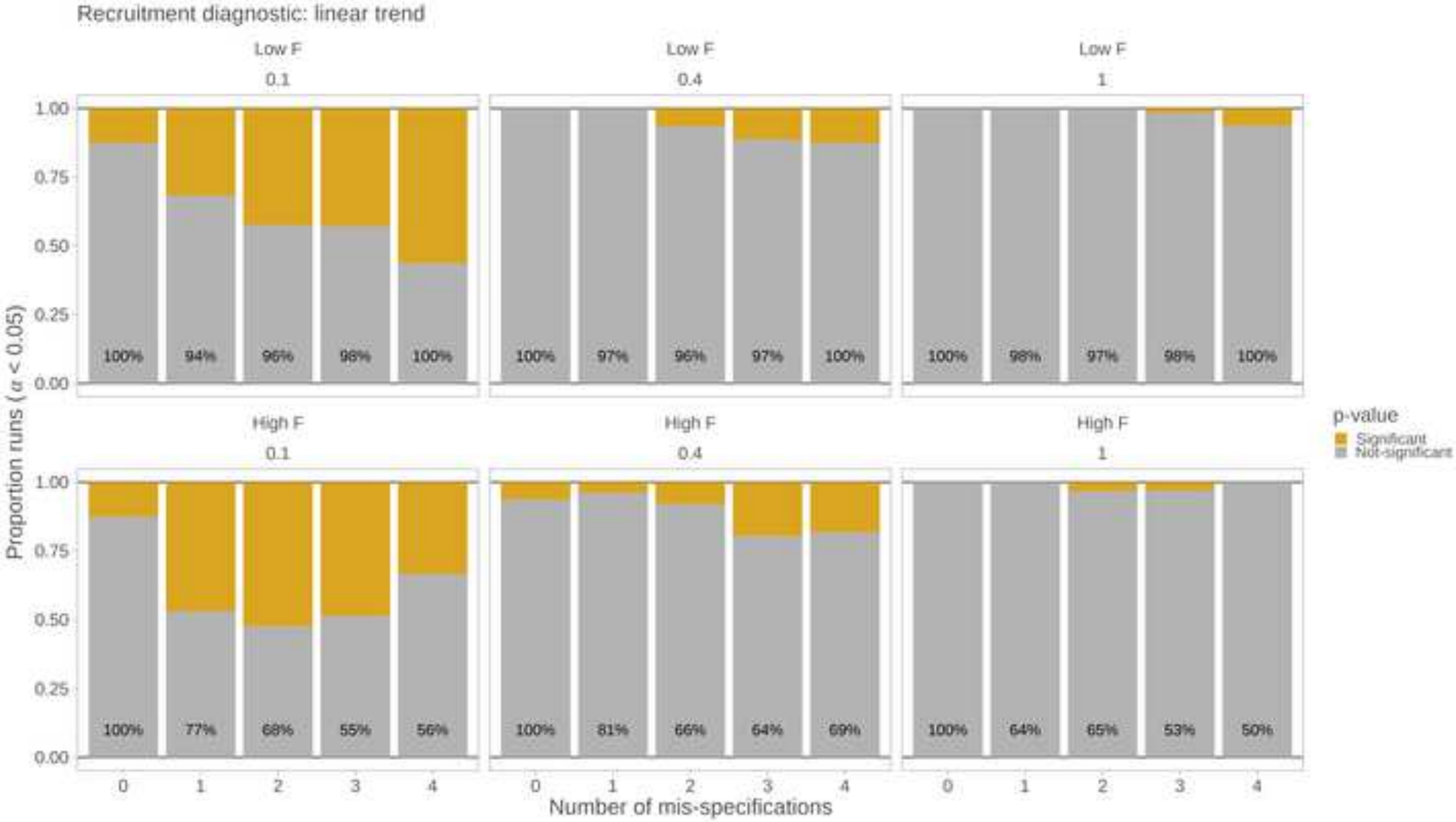



Spawning Stock Biomass
Fishing Mortality
Mohn's Rho
0.4
0.2
0.0
-0.2
Low F
High F
0
1
2
3
4
No. Misspecifications
sigmaR
0.1
0.4
1


sigmaR = 0.1
sigmaR = 0.4
sigmaR = 1
Low F
High F
Spawning stock biomass (1000 mt)
1.6e+07
1.2e+07
8.0e+06
4.0e+06
0.0e+00
0
25
50
75
100
Year
Model
Base
ASPM


R0 (RE)
MSY (RE)
SSB (MARE)
Low F
High F
Error
2
0
-2
0
1
2
3
4
No. Misspecifications
sigmaR
0.1
0.4
1


A) Survey
CPUE
Length
Age
Low F
High F
MASE
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
No. Misspecifications
B) Fishery
Length
Age
sigmaR
0.1
0.4
1


Recruitment diagnostic: linear trend
Low F
0.1
Low F
0.4
Low F
1
High F
0.1
High F
0.4
High F
1
Proportion runs (α < 0.05)
1.00
0.75
0.50
0.25
0.00
100%
94%
96%
98%
100%
100%
97%
96%
97%
100%
100%
98%
97%
98%
100%
100%
77%
68%
55%
56%
100%
81%
66%
64%
69%
100%
64%
65%
53%
50%
0
1
2
3
4
Number of mis-specifications
p-value
Significant
Not-significant