Model diagnostics: what are they good for?

This is a rough draft - final results forthcoming!

Maia S. Kapur[1], Nicholas Ducharme-Barth[2], Megumi Oshima[2], Henning Winker[3], and Felipe Carvalho[2].

[1]University of Washington, School of Aquatic and Fishery Sciences, Seattle, WA, United States
[2]NOAA Fisheries, Pacific Islands Fisheries Science Center, Honolulu, HI, United States
[3]Joint Research Centre (JRC), European Commission, TP 051, Via Enrico Fermi 2749, 21027 Ispra (VA), Italy

Keywords: Fisheries Assessment; Integrated Analysis; Model Diagnostics

Corresponding author 1: Felipe Carvalho
email: felipe.carvalho@noaa.gov
phone: 808 725 5605

Corresponding author 2: Maia S. Kapur
email: kapurm@uw.edu
phone: 206 526 4688

Abstract

Carvalho et al. (2021) provided a “cookbook” for implementing contemporary model diagnostics, which included convergence checks, examinations of fit, retrospectives and hindcasting, likelihood profiling, and model-free validation. However, it remains unclear how consistently such diagnostics can be employed to determine the nature and extent of model misspecification. In this study we use a state-space statistical catch-at-age simulation framework to compare diagnostic performance across a spectrum of correctly specified and mis-specified assessment models. We then contextualize 1) how reliably various diagnostic tests perform given the degree and nature of known model issues, including parameter, model process, and data misspecification, and combinations thereof; 2) how well diagnostic performance agrees with the direction and magnitude of mean average relative error in estimates of population size, and management quantities (reference points). We found that a surprising number of mis-specified models were indeed able to pass certain diagnostic tests. However, nearly all models which failed multiple tests were mis-specified, indicating the value of examining multiple diagnostics during model evaluation. 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.

Introduction

With the increase of computer power and the popularization of integrated stock assessment modeling (Maunder et al., 2009), the complexity of modern stock assessment modeling has rapidly increased as well. Analysts are typically confronted with different options to model the data and interpret the results when developing a stock assessment. Consequently, there has been surging interest in developing tools and suites for diagnosing and better understanding model performance. In this context, it is important to be specific about what a good model is. Carvalho et al. (2021) provided a "cookbook" for implementing modern model diagnostics and model performance evaluation. According to the "cookbook," a model would be considered adequate for providing management advice if the optimization was successful, the model fits the data (e.g., residual analysis), the model provides reliable estimates of trends and scale, the results of the model are consistent when updated with new data (e.g. retrospective analysis), and the model can make adequate future predictions (e.g. hindcasting).

In a perfect world, the suite of diagnostic tests proposed in the "cookbook" would be used to accept or reject candidate models leaving the "best" model remaining or to select a set of models to include in an ensemble. However, such an approach is still challenging as passing a diagnostic test, or a suite of tests cannot guarantee that the model adequately represents the "true" population dynamics, and failing one or more diagnostics, might be an indication that the model is mis-specified. As a result, this often leads to a key challenge for stock assessment modeling in practice, namely how one selects a model based on diagnostic results.

It is expected that models must be validated at some level if they are to provide credible and robust advice, and validation includes transferring confidence to people not directly involved in model construction. But ultimately, what is wanted is the validation of quantities of management interest, such as current Biomass and Fishing mortality. Still, we can't do that because such quantities are not observable. Therefore, we are essentially extrapolating outside the range of the data to latent quantities for providing management advice. This implies that our model assumptions must be approximately correct for the extrapolation to be helpful. For example, hindcast cross-validation of data unknown to the model is useful for validating the model's prediction skill of trends abundance. However, this approach can still not guarantee that the model correctly classifies the stock status and that the management advice is reliable. Therefore, it becomes critical to evaluate the diagnostics in terms of how well they diagnose the quantities of management interest rather than the model itself. If the diagnostics are used for weighting models, the weight might be based on the risk related to the bias in the management quantity of interest.

Given the potential impact that model diagnostics can have on scientists' and managers' fundamental understanding of contemporary models, this paper asks whether the diagnostics proposed in the "cookbook" are reliable and robust and to what extent. This study uses a state-space statistical catch-at-age simulation framework to compare diagnostic performance across a spectrum of correctly specified and mis-specified assessment models. We then contextualize 1) how reliably various diagnostic tests perform given the degree and nature of known model issues, including parameter, model process, and data misspecification, and combinations thereof; and 2) how well diagnostic performance agrees with the direction and magnitude of mean average relative error in estimates of population size, and management quantities (e.g., reference points).

Materials and Methods

Overview

We use a simulation approach to introduce a variety of misspecifications into the estimation model, and to evaluate the estimation performance in terms of relative error, diagnostic tests, and management quantities. The simulation procedure first involves the specification of a model of true population dynamics, the operating model (OM). 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 for both the fishery and surveys). All three types of data are generated from the OM for “1980-2016”, with a period of early recruitment deviations extending for 10 years prior to “1980” (Supplementary Figure S1). These generated data are used in a set of estimation models (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.

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 given in Table 1, and some biological assumptions (e.g., stock-recruitment steepness, natural mortality, and growth) were originally estimated for Pacific Hake (*Merluccius productus*, Taylor et al., 2015). A general description of the OM is as follows: it is a one-area, single-sex model, with time-invariant length-weight and maturity-at-age relationships, natural mortality (M), and length-at-age. 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) fixed at 0.86 and randomly-generated stock-recruitment deviations with a standard deviation (σ_R) of 1.4 in log space. The observation process involves a single fishing fleet and survey fleet. For the fishery fleet and survey, the relative probability of capture by age and length (selectivity) is time-invariant. Selectivity at length for the fishery fleet was parameterized to effectively be asymptotic. Selectivity for the fishery fleet was also assumed to vary by age. Selectivity at age 0 was assumed to be zero and selectivity at age 3+ was assumed to be 1. Selectivity at age 1 and 2 were freely estimated with the constraint that selectivity at age 2 be greater than age 1. For the survey fleet, all ages are assumed to be available to the survey with the same level of age selectivity. The initial conditions were specified so that there was no impact of fishing prior to the first year. Process error in each OM replicate arises from sampling a vector of recruitment deviates from a normal distribution with mean zero and variance of 1.4 (Supplementary Figure S1). This approach retains the general trend of the biomass time series across replicates while assuming the catches are known exactly and do not vary across replicates. This enables us to examine the impacts of mis-specified model components given an assumed population trajectory (with uncertainty), and to investigate the impacts of underreporting in catch.

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-composition data that provide a measure of the size structure of the index in the fishery. The catch observations were assumed to be known without error, as is reasonable for well-monitored fisheries. 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 0.1. Each length-composition observation was generated from a multinomial distribution with variability described by an effective sample size of 50. The values for the parameters for data generation follow the Hake-like estimation model of Lee et al. (2019, EM1). Below, we describe the model components that were manipulated in our simulation experiments and how the misspecifications were implemented.

Mis-specified processes

Growth

The growth curve in the OM is modeled using the von Bertalanffy (1938) 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 ($L\infty$, the inferred length at infinite age) and the growth rate K (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 model all main growth parameters ($L1$, L_∞, K, and CVs of length at ages) were freely estimated. For estimation models exploring a mis-specification in growth $L1$, L_∞, and K were held fixed at mis-specified values while CVs of length at ages remained estimable. Estimation models with L_∞ mis-specified are denoted by the letter L.

Natural mortality

Natural mortality, the instantaneous loss of population numbers due to causes other than fishing. In the OM, M is time- and age-invariant and set at 0.2 yr-1 (Table 1). Generally, it is difficult to obtain empirical estimates of this parameter for any fish species (e.g., Hamel, 2014). In fisheries, several methods have been developed that infer this value from the maximum age or length (Then et al., 2015) or via a meta-analysis of similar species within a genus (Thorson et al., 2017). Estimation models with natural mortality mis-specified at incorrect values are denoted by the letter M.

Reproduction

The regenerative capacity of a population influences its ability to recover from exploitation. In fisheries, the common measure of stock resilience is the steepness of the stock-recruitment relationship. Annual reproduction R in the OM is calculated based on a Beverton-Holt function (Equation 3) 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 an error term (set at 1.4), measure the log-distance from the deterministic curve given in Equation 1 and is a source of process error in the OM. Therefore, these recruitment deviates are randomly generated once for each OM replicate to reflect the process variability, but the steepness and R_0 are fixed (Table 1). In the estimation models, the recruitment deviates and R_0 are estimated with steepness fixed at either the correct or a mis-specified value. Estimation models with steepness mis-specified are denoted by the letter H (Table 1).

Fishery selectivity

In fisheries assessment models, "selectivity" refers to the proportional length- or age-based availability of fish to surveys or fisheries. In the OM, the fishery uses a length-based asymptotic selectivity pattern, meaning that all individuals above a certain size have a close to equal probability of being captured (Table 1). When the selectivity is correct, estimation models estimate the selectivity parameters under the assumption that selectivity is an asymptotic function of length for the fishery. Estimation models with selectivity mis-specified are denoted by the letter X, indicating that the model fixes the length of fully selected individuals at a mis-specified value.

Determining misspecification thresholds

Mis-specified parameter values were fixed at +/- 15% of the correctly specified values. The exception to this was for the L_∞ parameter which was restricted to +5% given the maximum assumed population length bin in the model. . Values were selected based on conducting univariate parameter sweeps of the mis-specified parameters using a simplified model. The simplified model only estimated LN(R_0) and held all parameters fixed at correct values except for the target mis-specification which was held fixed at a mis-specified value (e.g., +/- 75% of the correct value at 1% intervals). Applying this procedure allowed us to determine the amount of error introduced into terminal estimates of SSB, F, and B/B_0 by each mis-specification. Mis-specifying parameter values by +/- 15% was found to generally result in ~10% - 20% relative error in terminal estimates using the simple model. This verification ensured that the misspecifications implemented in our experimental design are known to impact estimated outputs to a similar extent. This led to two mis-specified parameter values, one above and one below the original values used in the OM, for all parameters.

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 (Figures 1 and 2), which enabled to determine how well model diagnostics could detect the nature and extent of model misspecification. 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 three values for each replicate, each with an even probability of being either a 0 or 1. This vector determines how each mis-specification is implemented. For example, the first OM replicate may have the draw [0, 0, 0] in which all three 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 functionally unique 16 "strings" corresponding to the mis-specified categories (Table 1) to each replicate. All unique combinations of mis-specifications were evaluated. For example the string "XMLH" denotes a model with all four mis-specifications, while "XM" denotes a model with only selectivity (X) and natural mortality (M) mis-specified. Note that "XM" is functionally equivalent to "MX" 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 twenty-six 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 string. In total, the study design fit 416 EMs (16 unique estimation models fitted to 26 OM replicates).

Data Misspecification

To evaluate the impact of data misspecification on model performance, we explored a limited suite of data misspecifications, which included hyperstability and hyperdepletion of survey indices, and underreporting of catch. Hyperstability occurs when a CUPE index declines more slowly than the abundance and hyper depletion is the opposite scenario, when a CPUE index declines faster than the abundance. In both of these instances, the CPUE index is not accurately tracking the abundance of the stock which can lead to incorrect estimation of management quantities and result in lost yield or over exploitation. Another common data problem in many fisheries is underreporting of catch. Underreporting rates can decline over time if a fishery has increased observer coverage or improved estimates of discards. Conversely, underreporting rates can increase as a response to new management restrictions or if there is an increase in recreational fishing (Rudd and Branch 2017). All data misspecification experiments were conducted separately from the combinatory experimental design described above, so that all EMs with either data misspecification were otherwise correctly specified (M, steepness, growth and selectivity parameters set to the same behavior as in the OM).

For the hyperstability analysis, novel survey indices were constructed after the bootstrapping exercise following Erisman et al., 2011, using a range of beta values from 0.1 to 0.9 in increments of 0.1. The same vector of observation errors for years 1980-2016 was used for each hyperstable index. This approach represents a scenario wherein each survey year is subject to the same value of observation error, while the only difference among survey time series is the degree of hyperstability affecting the index's sensitivity to population change. We chose to do so to disentangle the effect of the hyperstable index on our performance measures from simple observation noise, which would have been the case if sigma were randomly drawn for each year beta combination in the series. Hyperdepletion was implemented following the same method, using a range of beta values from 1.1 to 2.0 in increments of 0.1.

For the catch underreporting analysis, catch time series were constructed after the bootstrapping exercise as:

$$C_y = C_y^{true} * R_y \qquad \text{Eq. 2}$$

where C_y^{true} is the true catch in year y and R_y is the underreporting rate in year y. We tested four underreporting scenarios, one where underreporting decreases over time, and three where underreporting increases over time. All scenarios allowed the reporting rate to change from the initially specified rate to the final specified rate according to a Brownian bridge. When underreporting decreased, catch was reported at 50% of the actual catch in the beginning of the time series and gradually increased until reaching 100% by the end year. When underreporting increased over time, catch reporting 1) decreased to 50% of actual catch over the entire time period, 2) decreased to 50% of actual catch starting after 1990 to the end of time series and 3) decreased to 50% of actual catch starting after 2000 to the end of the time series.

Computing

Calculation of model diagnostics across 416 EMs and 624 data mis-specifications 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).

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*=1980,1981…2016), replicate (*j*=1,2,…24), string (*k*=1,2…6), and model (*i*=1,2,…26) were summarized using relative (equation 3) or absolute relative errors (equation 4) and then averaged across replicates.

$${}_{SSB_y}^{EM_{i,j,k}} = \frac{\sum_i \frac{\widehat{SSB}_y^{EM_{i,j,k}} - SSB_y^{OM_i}}{SSB_y^{OM_i}}}{26} \qquad \text{Eq. 3}$$

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 SSB are calculated via:

$$SSB_{y=2007-2016}^{EM_{i,j,k}} = \frac{\sum_i \frac{\left| \sum_{y=2007}^{2016} \frac{\widehat{SSB}_y^{EM_{i,j,k}}}{10} - \sum_{y=2007}^{2016} \frac{SSB_y^{OM_i}}{10} \right|}{\sum_{y=2007}^{2016} \frac{SSB_y^{OM_i}}{10}}}{26} \qquad \text{Eq. 4}$$

Model Diagnostics

Model diagnostics were developed following the recommendations of the cookbook (Carvalho et al., 2021) and using the R package ‘ss3diags’.

Residual analysis

Patterns in residuals (i.e. presence of nonrandom variation) were explored using a non-parametric runs test. Passing of failing the run-test is judged by the p-values computed for each time-series, where $p \geq 0.05$ suggesting no evidence to reject the hypothesis of randomly distributed residuals. In addition, outlier data points were identified via the 3-sigma limit, where any points beyond 3 SD would be unlikely, given random process error in the observed residual distribution. Any extreme patterns of positive or negative residuals are indicative of poor model performance and potential unaccounted for process or observation error. The third measure used in the residual analysis was the root mean square error (RMSE), which was calculated for overall model fit of the relative abundance indices and composition data. A relatively small RMSE ($\leq$ 0.3) indicates a reasonably precise model fit to relative abundance indices (Winker et al., 2018).

Likelihood Profiles

Profile likelihoods are used to examine the change in log-likelihood for each data source in order to address the stability of a given parameter estimate, and to see how each individual data source influences the estimate. When a given parameter is not well estimated, the profile plot may show conflicting signals across the data sources. The resulting total likelihood surface will often be flat, indicating that multiple parameter values are equally likely given the data. In such instances, the model assumptions need to be reconsidered.

We constructed likelihood profiles on R_0 using the SS_doProfile function from R package r4ss for each estimation model (Taylor et al., 2011). This approach sequentially fixes unfished recruitment at a pre-specified value and re-runs the estimation model with whatever other parameter settings were specified in the original experiment. This was repeated for all of the unique OM replicate-estimation model combinations described previously. The range of R0 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 model associated with that replicate, in increments of 0.01. As part of the exploration of this diagnostic, we calculated the ψ statistic recommended by Wang et. al.,

$$\varphi = \begin{cases} \max\left[\left(L^{c}_{lower} - L^{MLE}\right), \left(L^{c}_{upper} - L^{MLE}\right)\right], \text{if } R^{c}_{0} \text{ is located within } 95\%\ CI \text{ for } R^{MLE}_{0} \\ \left|L^{c}_{lower} - L^{c}_{upper}\right|, otherwise \end{cases} \qquad \text{Eq. 5}$$

where ψ is a 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). For simplicity, we focus on the Survey and Length Composition data.

Retrospectives

A retrospective analysis is a useful approach for addressing the consistency of terminal year model estimates. The analysis sequentially removes a year of data at a time and reruns the model. If the resulting estimates of derived quantities such as SSB differ significantly, particularly if there is serial over- or underestimation of any important quantities, it can indicate that the model has some unidentified process error, and requires reassessing model assumptions.

ASPM & Deterministic recruitment model

Maunder and Piner (2015) proposed an age-structured production model (ASPM) as a model diagnostic for complex age-structured integrated assessments. The application of an ASPM diagnostic can detect misspecification of key systems-modeled processes that control the shape of the production function (Carvalho et al., 2017). The ASPM performs well in a system with informative contrast. A system with informative contrast likely means that fishing effort is high and measurements from the system (i.e., catch, life history, and index) are reasonable representations of the actual states. If the ASPM fits well to the indices of abundance the production function is likely to drive the stock dynamics and the indices will provide information about absolute abundance (Minte-Vera et al., 2017). On the other hand, if there is not a good fit to the indices, then the catch data and the production function alone cannot explain the trajectories depicted in the indices of relative abundance. This could be due to mis-specification of the components which make up the production function. 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 another way that can be used to diagnose the models ability to capture the production function, and 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, the strength of the production function was measured by calculating the relative difference in model estimates of R0, MSY, and the mean absolute difference (MARE) in predicted SSB between the full model and the ASPM/deterministic recruitment model.

Hindcast Cross Validation

Accuracy and precision of a model's prediction skill can be measured with hindcast cross validation which involves comparing observations to predicted future values. 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 with the removed observations and prediction skill can be measured by the mean absolute scaled error (MASE). A model predicts better than a random walk (naive model) if the MASE is less than 1; if MASE is greater than 1, the model predicts worse than a random walk model. MASE can be applied to index of abundance and compositional data.

Recruitment trend

It is generally recognized that patterns in residuals is indicative of model mis-specification and un-modeled process. The estimation of recruitment deviates is one of the principle ways that process error is incorporated in stock assessments, so trend in the recruitment deviates could be an indication of model mis-specification (Merino et al. 2022). Following Merino et al. 2022 we quantified whether a significant linear trend in the recruitment deviates existed. We also tested for monotonic trends, and non-monotonic trends in recruitment deviates. Additionally, we calculated the degree of first order temporal autocorrelation in the deviates and applied runs tests.

Results

Summary of Simulations

For the main (combinatory) experiment, a total of 416 models were run. All models were found to be converged with final gradients smaller than 5.1E-04 (median 6.3E-06). For the hyperstability and hyperdepletion analyses, a total of 520 models were run and all models reached convergence. Overall, as beta values increased or decreased farther away from one, model diagnostic performance decreased and hyperdepletion had a greater effect on model diagnostic performance than hyperstability. For catch underreporting analyses, we ran a total of 104 models and all reached convergence. Overall, model diagnostics were not very sensitive to catch underreporting, with the majority of model runs passing all diagnostics.

Relative Error

In the main experiment, MARE in terminal SSB ranged from an average of 5% to 18%, increasing with the number of misspecifications present. MARE in terminal F similarly increased from 4% from the correct estimation model to 8% in models with four misspecifications. Relative error of terminal SSB and F were both sensitive to hyperstable and hyperdepleted CPUE indices. Relative error ranged from -0.91 to 4.99 for terminal SSB and -0.75 to 5.74 for terminal F. Generally, when beta was less than one (hyperstable) SSB was overestimated and F was underestimated. When beta was greater than one (hyperdepleted), the opposite trend was observed, SSB was underestimated and F was overestimated. The magnitude at which F or SSB was incorrectly estimated varied widely between and among OM replicates. Relative error of terminal SSB and F were sensitive to catch underreporting. Relative error ranged from 0.09 to 0.97 for terminal SSB and -0.39 to 0.06 for terminal F. For all cases of underreporting except one, F was underestimated from the true value and SSB was overestimated for all cases, but the magnitude of relative error varied across OM replicates.

Residuals

The residual analyses included examination of the normality of survey residuals, as well as the calculation of the root-mean-square error (RMSE) for both the length composition and survey biomass data. There were clear trends of increasing RMSE for both data types with an increasing number of misspecifications (Figure 3). Generally, models which had mis-specified growth parameters had the highest mean RMSE for both data types, followed by those with mis-specified selectivity. Results for the run's test in the main simulation were more mixed: while 95% of "correct" estimation models passed the runs test for both data types, no more than 30% of highly mis-specified models failed.

The runs test was very sensitive to the level of stability or depletion of the survey CPUE index. As the CPUE index became more stable (beta < 1), the proportion of runs test that failed increased until once beta = 0.2 all replicates failed. A similar pattern occurred for hyper depletion, the proportion of runs test that failed increased as beta increases closer to 2.0. The length composition data experienced the same trend but at a much smaller scale. The maximum proportion of runs test that failed were 27% and 23% for the fishery and survey data respectively.

RMSE of the CPUE index increased as beta moved away from 1 in either direction, however it was greater when beta was greater than one, or when hyper depletion was occurring. In some instances, it was over 90% (Figure S2). RMSE of length composition data showed a similar trend with RMSE increasing as beta moves farther from 1 but at a much smaller magnitude and the median RMSE remained much lower and fairly consistent across beta values.

Overall, the runs test was fairly insensitive to catch underreporting. Length composition and CPUE data from the survey fleet had the greatest proportion of failed tests across the underreporting scenarios. When underreporting decreased over time, 26% of the runs test failed for the survey fleet length composition data. When underreporting increased over the entire time period, 23% of the runs test failed for the CPUE indices. RMSE was much lower for the fishery length composition data than the survey length composition data or CPUE indices for all catch underreporting scenarios. There was also more variability in RMSE for survey length composition data and CPUE indices than for the fishery length data.

Likelihood profile

Of the 83,616 unique models run as part of the profiling exercise, three did not converge and were excluded from this analysis. Figure 4 shows a simplified illustration of likelihood profiles by component, which are scaled so that the x-axis represents the difference between the fixed R0 for the model at hand and the value for R0 from the OM (which did not change across replicates). There was some spread in the MLE of R0 indicated by the total likelihood, though all profiles were well-defined. For all correct EMs, profiles for the survey data were skewed left, suggesting higher MLEs for R0 but with lower specificity (many statistically indistinguishable models below the MLE). Additionally, the length composition data was generally informative, with well-defined profiles that were roughly centered just at or above the R0 from the OM. The MLEs from the length composition data in correct EMs were slightly lower than the overall MLEs.

Upon introducing misspecifications, the likelihood profiles changed systematically from those obtained using the correct EM. The overall MLE for R0 drifts positively with sequential inclusion of misspecifications, resulting in a most discrepant MLE 0.4 units greater (in log space) than the correct MLE. The pattern of rightward drift is consistent across both data components, whereby greater degrees of misspecification result in higher estimates of R0 relative to the "correct" model. However, the qualitative and relative behavior of the profiles was strikingly consistent within EMs: the Survey data were always the broadest, with MLEs above the total, and the length composition profiles were consistently narrower than the Survey and shifted slightly below the total MLEs.

The ψ statistic provided a metric of the information content of each data component with the sequential introduction of misspecifications. The correct EMs estimated the length composition data (ψ=0.0650) to be consistently more informative than the survey data (ψ=−0.0456), with no variation among OM replicates. With one or more misspecifications, the spread of ψ values increased for both data types, with a nontrivial proportion of mis-specified models suggesting the length-composition data were more informative when the model was indeed mis-specified. Indeed, ψ statistics for all fully-mis-specified models suggested that the length composition data were more informative than the correct EMs (ψ=0.0750), though they consistently indicated the least information obtainable from the survey data (ψ=−0.0929).

For the data misspecifications, of the 124,620 unique model runs for the profiling exercise, 276 did not converge and were excluded from the analysis. When the survey data was mis-specified, hyperstability (beta < 1) had a greater impact on the composition data than on the survey data. While the MLE of R0 suggested by the composition data was always lower, the MLE of R0 suggested by the survey data was closer to the overall MLE and was only shifted positively once beta was less than 0.5. In contrast, hyperdepletion (beta > 1) had a greater impact on the survey data component (Figure S3). The MLE of R0 was slightly negative relative to the overall MLE but the survey data showed a strong right skew with low specificity and the MLE of R0 was suggested to be 1.0 units lower (in log space) than the overall MLE. For all instances when CPUE indices were hyperstable, the length composition data were estimated to be more informative (ψ ranged from 0.43 to 0.09) than the survey data (ψ ranged from -0.01 to -0.06). However, when CPUE indices showed hyperdepletion, survey data were estimated to be more informative than the length composition data. At the maximum level of hyperdepletion (beta = 2), ψ statistics suggested that the survey data was more informative (ψ = 0.14) than the length composition data (ψ = 0.01).

For the catch underreporting misspecification, the MLE of R0 suggested by the survey data was always lower than the overall MLE of R0 and not well defined (Figure S3). However, the MLE suggested by the composition data was only slightly lower than the overall MLE and well defined. The ψ statistic suggests that the length composition data was more informative than the survey data for every catch underreporting scenario.

Retrospectives

There was not a statistically significant trend in mohn's rho with increasing levels of misspecification. All correct estimation models fell within the acceptable range of -0.15 to 0.2 for SSB and F, while 12% of completely mis-specified models had values for rho for biomass outside of this range (Figure 5).

Mohn's rho was not very sensitive to hyperstability for either SSB or F, but was very sensitive to hyperdepletion for both SSB and F (Figure S4). For SSB, a small number of runs when beta was less than 1 fell outside of the suggested range of -0.15 to 0.2 but the median Mohn's rho was very close to zero for all hyperstable runs. Conversely, as beta increased from one, the median Mohn's rho values decreased and for all the beta values tested greater than 1.5 the median Mohn's rho was below the lower threshold. Additionally, the spread of values increased as beta increased. For F, beta values less than one exhibited a similar pattern to SSB, with a slightly higher median value but only a small number of runs were outside of the suggested range. As beta increased greater than one, the median and spread of Mohn's rho values increased and for any beta values greater than 1.6, the median Mohn's rho value fell above the upper threshold.

Mohn's rho statistic was not sensitive to catch underreporting for either SSB or F. While there was a small bias in both, none of the Mohn's Rho values fell outside of the window of -0.15 and 0.2 suggested by Hurtado-Ferro et al (2015). SSB exhibited positive bias for scenarios where underreporting increased over time and negative bias for the scenario where underreporting decreased over time. F exhibited the opposite behavior, it was negatively biased for the scenarios with increasing underreporting over time and positively biased for the scenario with decreased underreporting.

ASPM

Both the ASPM and deterministic recruitment results showed virtually identical patterns in terms of differences in R0 between the two models, and the MARE of SSB (Figure 6). Median (across models) of the relative difference in R0 was greatest in the correctly specified model, and as the number of mis-specifications increased the relative difference increased. This pattern was mirrored for the difference in MSY estimates for the ASPM, though for the deterministic recruitment model the median difference stayed close to 0 as the number of mis-specifications increased. For the MARE of SSB, the MARE increased with the number of mis-specifications for both the ASPM and the deterministic recruitment model. Despite some trend in metric medians as the number of mis-specifications increased, there was considerable overlap in the interquartile range for all cases.

For the data misspecifications, the ASPM and deterministic recruitment results showed almost identical patterns across the three metrics (Figure S5). For relative error of R0 for both ASPM and deterministic recruitment, as hyperstability increased, the median increased and as hyperdepletion increased the relative difference in R0 decreased. Relative error of MSY for both diagnostics was higher for models with greater hyperstability but as CPUE indices became less hyperstable and more hyperdepleted, the relative error decreased. For the ASPM diagnostic, extreme hyperdepletion (beta $\geq$ 1.8) led to median relative error very close to zero. For deterministic recruitment, small levels of hyperstability (0.6 $\leq$ beta $\leq$ 0.8) led to median relative error very close to zero as well but as beta increased or decreased outside of the range, median relative errors moved farther away from the correct model. The opposite trend occurred for MARE of SSB for both ASPM and deterministic recruitment. As hyperstability increased, the median MARE of SSB decreased relative to the correct model but as hyperdepletion increased, the median increased. For the catch underreporting misspecifications, for both ASPM and deterministic recruitment diagnostics, median relative error of R0 and MSY was highest when catch was underreported early in the time series but median MARE of SSB was lowest (Figure S5). For all three scenarios with catch underreporting increasing over time the median relative error of MSY was close to zero for the deterministic recruitment diagnostic.

Hindcast

Hindcast cross-validation MASE scores for the CPUE index, survey mean length composition, and fishery mean length composition tended to increase as a function of the number of model mis-specifications indicating worse performance (Figure 7). However, this deteriorating performance in terms of hindcasting was most notable for the fishery mean length composition. When considering hindcast predictive performance relative to the null model, a majority of all models in each mis-specification class indicated predictive performance greater than the null model (e.g. MASE < 1) each for CPUE, survey mean length, and fishery mean length. Models that simultaneously mis-specified both growth and selectivity had noticeably poorer MASE scores for the fishery mean length composition relative to other models.

Hindcast cross-validation prediction skill was sensitive for CPUE indices for both hyperstable and hyperdepletion scenarios, however, for length composition, only hyperdepletion had a large impact. For CPUE indices, as beta moves away from one, prediction skill of the model decreases (MASE increases) dramatically (Figure S6). For an extremely hyperstable index, the median MASE score was 1.75 and the max was 7.63, indicating the model was almost eight times worse at predicting CPUE than a naïve model. Prediction skill of length composition data was less sensitive to the hyperstable CPUE index, however, it was sensitive to the hyperdepleted CPUE, particularly for the survey length composition data. For both fleets, prediction skill decreased as beta increased towards 2.

Catch misspecification most impacted the prediction skill of the model for length composition data from the survey fleet, and least impacted the prediction skill of the model for length composition data from the fishery (Figure S6). For all three scenarios with underreporting rate increasing over time, there were 28 models with MASE score greater than one for survey length composition. However, the median MASE score was below one for all scenarios. Conversely, MASE scores for the fishery length composition data was greater than one for only three models across all four underreporting scenarios. Median MASE scores were 0.5 for decreasing underreporting, and 0.6 for all three increasing underreporting scenarios which can be interpreted as the models are nearly twice as good at predicting fishery length composition data as a naïve model is. Median MASE scores for the CPUE indices were all below one.

Recruitment

The recruitment trend diagnostics were largely inconclusive as close to 100% of all models did not indicate linear trend, monotonic trend or any trend in recruitment deviates (Figure 8). Additionally, close to 100% of all models did not show signs of first order temporal autocorrelation in the recruitment deviates, or sequences of runs in the residuals. Similarly, for the data misspecifications, the recruitment trend diagnostics were largely inconclusive (Figure S7). Nearly all models for hyperstable CPUE, hyperdepleted CPUE, and catch underreporting did not indicated linear, monotonic, or any trend in recruitment deviates. Almost all models showed no first order temporal autocorrelation or sequences of runs as well.

Discussion

Performance of individual diagnostics

RMSE/Residuals seem to logically respond to misspecifications in processes that degrade survey fits (growth and selectivity). However, runs test results were more robust in the context of hyperstability (with other model processes correctly specified). A majority of highly-mis-specified models proceeded to pass the runs test, which suggests that either the model(s) were able to sufficiently compensate in terms of survey fits, or that the statistical cutoff for passing the test is not appropriate. If the former is true, it is unlikely that the analyst would be dissatisfied with the survey fits to begin with, and would likely need to examine other model results. Out of all diagnostics, the RMSE test most reliably returned higher values with increasing amount of misspecifications; the runs test also regularly failed when the cpue series was either hyperstable or hyperdepleted. This is reassuring evidence that goodness-of-fit tests can be a useful first step in evaluating a model.

Mohn’s rho calculated from retrospective analyses was a surprisingly poor correlate to model misspecification (given the traditional cutoffs range for rho of -0.15 to 0.2), though retrospective performance did degrade with increasing misspecifications. 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. Our study did not investigate misspecification of time-varying processes, which might induce stronger retrospective bias.

This study demonstrates the versatility of the ASPM and deterministic recruitment model diagnostics. Carvalho et al., (2017) showed via simulation analyzes that ASPM was the only tested diagnostic capable of detecting misspecification 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, basically confirming this example as a recruitment driven model. It is also important to note the deterministic recruitment model shows virtually the exact same results obtained from the ASPM, so this diagnostic can also be used as an alternative to measure the effects of fishing.
For most data components the value of MASE was closely related to the number of model misspecifications, which is an indication that the hindcast diagnostic provides a way to evaluate

model performance, and can potentially identify model misspecification. However, predictive skill, in the form of the MASE criterion, were satisfactory across all models. This is surprising, as in general, good MASE performance for CPUE hindcasting is likely to occur if the stock is production driven and the production function is estimable from the data (Maunder et al., 2022), which is not the case of our example.

Likelihood profiles appear to remain internally consistent, with the relative degree of conflict stable across mis-specified models (e.g. survey data were always less informative than 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.

Study Caveats & Major Takeaways

A surprising number of mis-specified models both fit the data reasonably well and were able to pass individual diagnostics (or did not show clear trends in worsening diagnostic performance given the presence of misspecification). However, there are several characteristics of our study design that should be considered alongside this result, and form the basis for additional studies regarding the utility and robustness of the diagnostic tools examined here.

All operating models used an identical catch series. Some diagnostics, like MASE, might simply echo the responsiveness to fishing pressure, which in this case will be more pronounced in trajectories that have lower SSB as a result of the recruitment curve. As stated in Punt et al. (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 (e.g. M). 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.

A holistic examination of our results produced several insights: firstly, that the OM structure is highly recruitment driven, such that virtually all estimation models were able to recover roughly the same recruitment trend as the operating model. This is likely because of the “data-richness” of the simulation, in that the length composition data (from which recruitment estimates are derived) is much more informative than the survey. In addition, there was clustering of diagnostic performance with low, and medium or high MARE in SSB, suggesting that diagnostic performance might be more robust only at the point where derived spawning biomass is well above or below the real value. Given that actual stock assessors are ignorant to “reality”, these results would mainly be useful in the case that other model results (such as fits to survey and compositional data) appeared satisfactory.

Good Practices in Applying Model Diagnostics

The primary challenge in developing diagnostic workflows arises because real stock assessors must evaluate a small subset of total possible models representing a population. In our study, the RMSE and likelihood profile diagnostics were the most internally consistent and responsive to the presence of misspecification. Yet an assessment scientist would only see one or two of these models' results, and have no knowledge of how divergent the result 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's likely that models with poor RMSE scores would have been discarded in the first place based on their poor fits to the survey data.

Overall, our results did not provide a very clear insight when looking at how diagnostics relate to MARE for SSB, and it is likely that the patterns that do emerge are closely related to the simulated population dynamics of the stock (i.e. OM driven). However, the performance of some diagnostics did vary based on the misspecifications. For example, models with low RMSE for the composition and CPUE data, and the low MASE scores for composition data showed the lowest SSB MARE.

At present, these results render it difficult to develop robust thresholds for diagnostic tools that can be applied across all stocks. Further investigation of diagnostic performance should evaluate 1) how dependent these results are upon the case study having well-informed recruitment, 2) relatedly, if there are correlations between the operating model characteristics (e.g. general stock trajectory) 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. To address the first and second topic, this project will add, prior to final submission, operating models that are production driven, have less informative compositional data, or perhaps use a lower value for recruitment variability (σR).

References

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/.

Erisman, B.E., Allen, L.G., Claisse, J.T., Pondella, D.J., Miller, E.F., Murray, J.H., and Walters, C.J. 2011. The illusion of plenty: hyperstability masks collapses in two recreational fisheries that target fish spawning aggregations. Canadian Journal of Fisheries and Aquatic Sciences 68(10): 1705–1716.

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.

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.
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.

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.

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

Rudd, M.B., Branch, T.A. 2017. Does unreported catch lead to overfishing? Fish and Fisheries. 18(2): 313–323.

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.

Taylor, I.G., Grandin, C., Hicks, A.C., Taylor, N. and Cox, S. 2015. Status of the Pacifc Hake (whiting) stock in U.S. and Canadian waters in 2015. Prepared by the Joint Technical Committee of the U.S. and Canada Pacifc Hake/Whiting Agreement; National Marine Fishery Service; Canada Department of Fisheries and Oceans. 159 p.

Tables

Table 1. Summary of experimental design, sources of uncertainty, treatment of mis-specified models, and number of models run in each category. The values in parentheses are the

	Mis-specified Parameters/Processes				**Data Inputs**			
Model Description	steepness	natural mortality	selectivity	growth	Catch data	Survey data	Length composition data	No. models/replicates
OM	0.85	0.2 yr-1	Fishery: logistic with inflection point at 39 cm	L2 = 53.8 cm,	Annual catches 1979-2017, CV = 0.01	Annual observations 1980-2016, CV = 0.1	Annual observations 1980-2016, 8-52cm bins, input sample size of 50	27
“Correct” EM	As in OM	As in OM	As in OM	As in OM	As in OM	As in bootstrapped replicate	As in bootstrapped replicate	27
EMs with *h* mis-specified	0.73 or 0.99	As in OM	As in OM	As in OM	“	“	“	208
EMs with *M* mis-specified	As in OM	0.17 or 0.23	As in OM	As in OM		“	“	208
EMs with selectivity mis-specified	As in OM	As in OM	33.4 cm or 45.1 cm	As in OM		“	“	208

EMs with *growth* mis-specified	As in OM	As in OM	As in OM	45.8 cm or 56.6 cm		“	“	208
Data Quality Experiments								
EMs with hyperstable survey series	As in OM	“				As in bootstrapped replicate, made more or less stable following Erisman et. al (2011)		
EMs with under-reported catch								

Figures

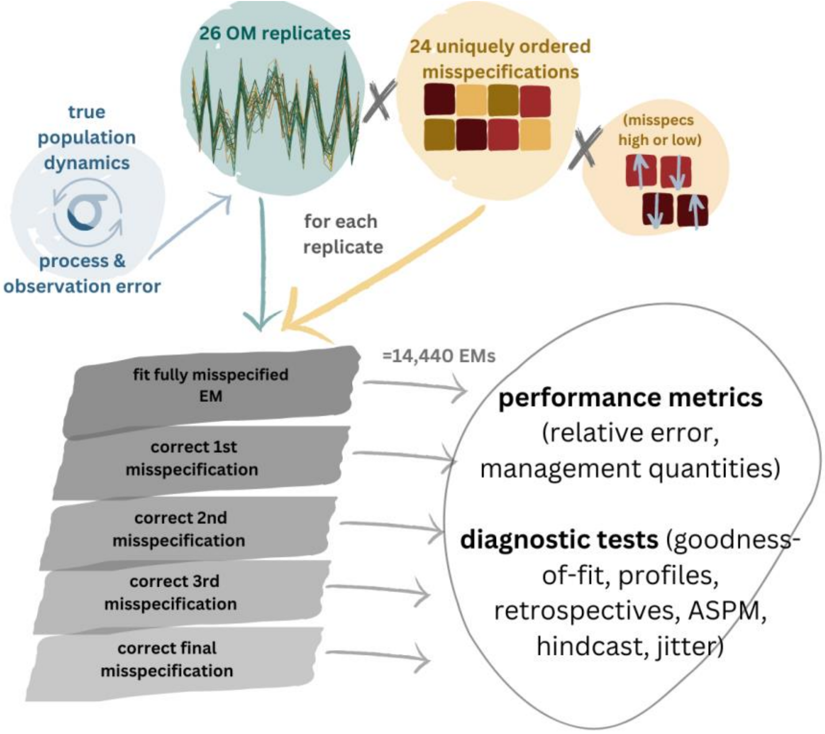

26 OM replicates
24 uniquely ordered misspecifications
(misspecs high or low)
true population dynamics
process & observation error
for each replicate
=14,440 EMs
fit fully misspecified EM
correct 1st misspecification
correct 2nd misspecification
correct 3rd misspecification
correct final misspecification
performance metrics (relative error, management quantities)
diagnostic tests (goodness-of-fit, profiles, retrospectives, ASPM, hindcast, jitter)

Figure 1. 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 models, we compute performance metrics (relative error with respect to the OM replicate at hand) and run a suite of diagnostic tests as described in Carvalho et al. (2021).

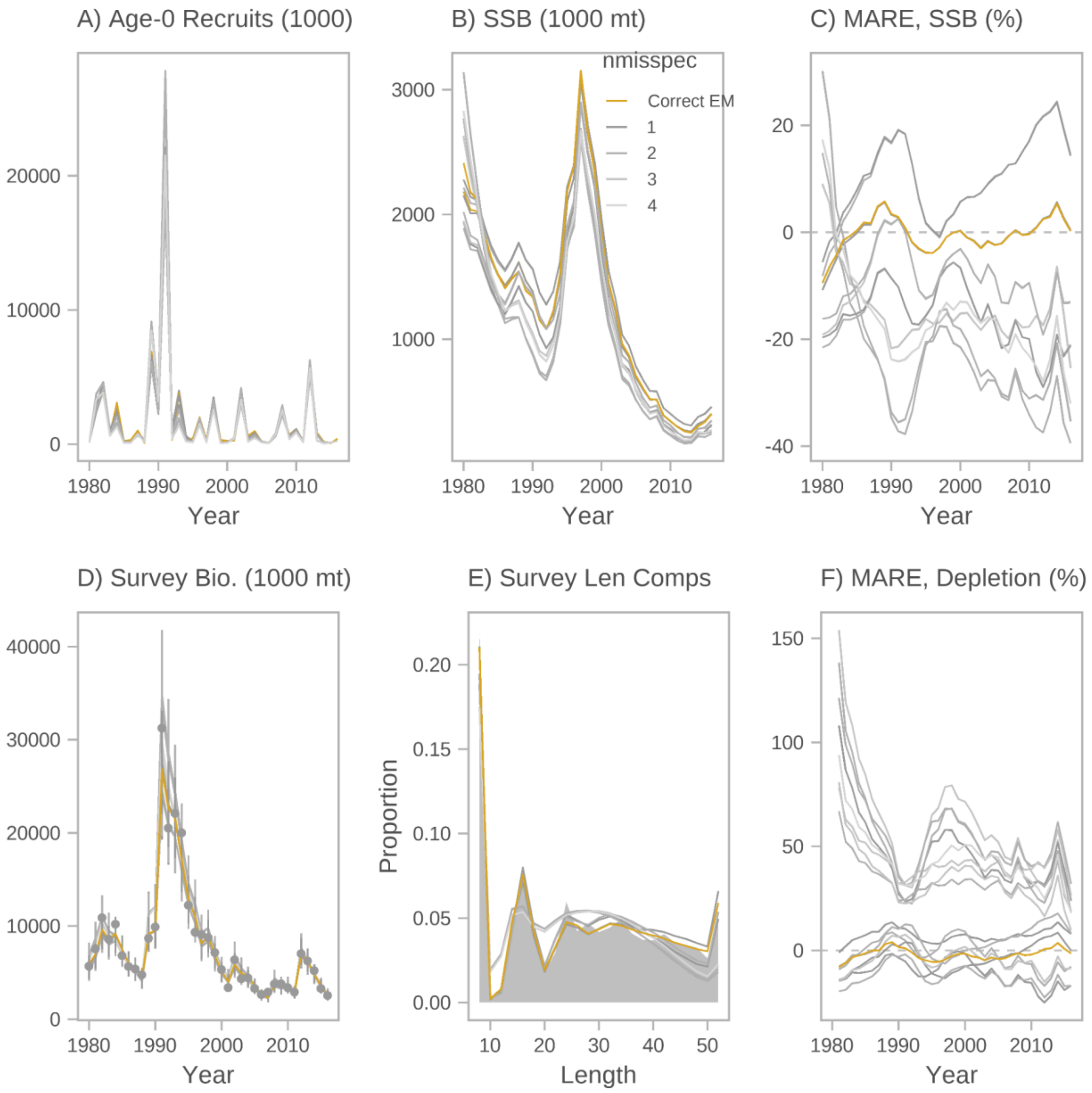


Figure 2. Illustration of single simulation replicate. Each panel shows trajectories from all estimation models fit to a single replicate of the OM, including A) age-0 recruits, B) stock spawning biomass, C) mean average relative error in SSB, D) expected (points) and observed (lines) survey biomass, E) aggregated length composition observations (polygon) and fits (lines), and F) mean average relative error in depletion. The line colors correspond to the number of misspecifications present in the estimation Model; the gold trajectory corresponds to the correct (un-mis-specified) estimation model).

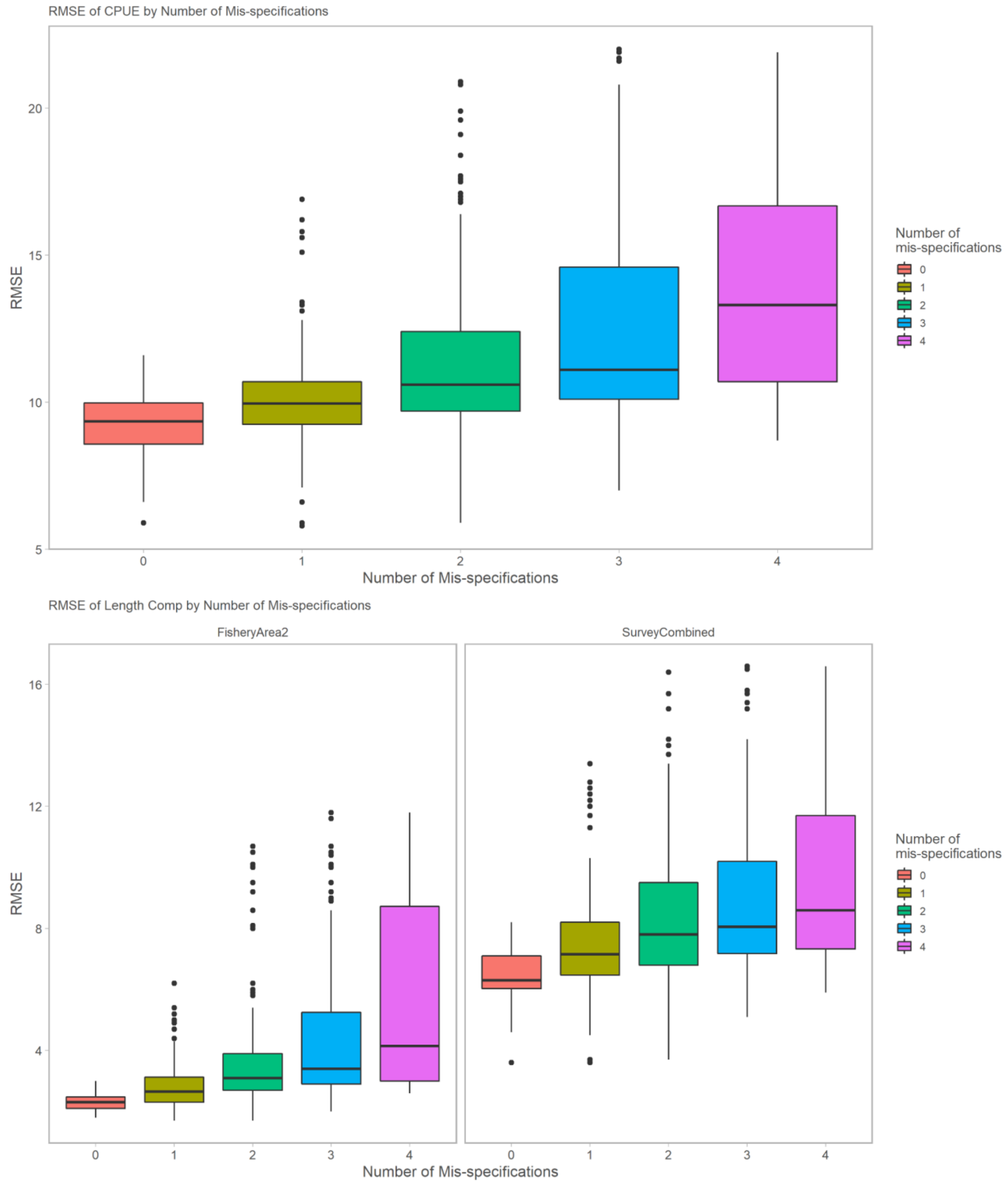


Figure 3. RMSE of Survey data (top) and length composition data (bottom), grouped by the number of misspecifications.

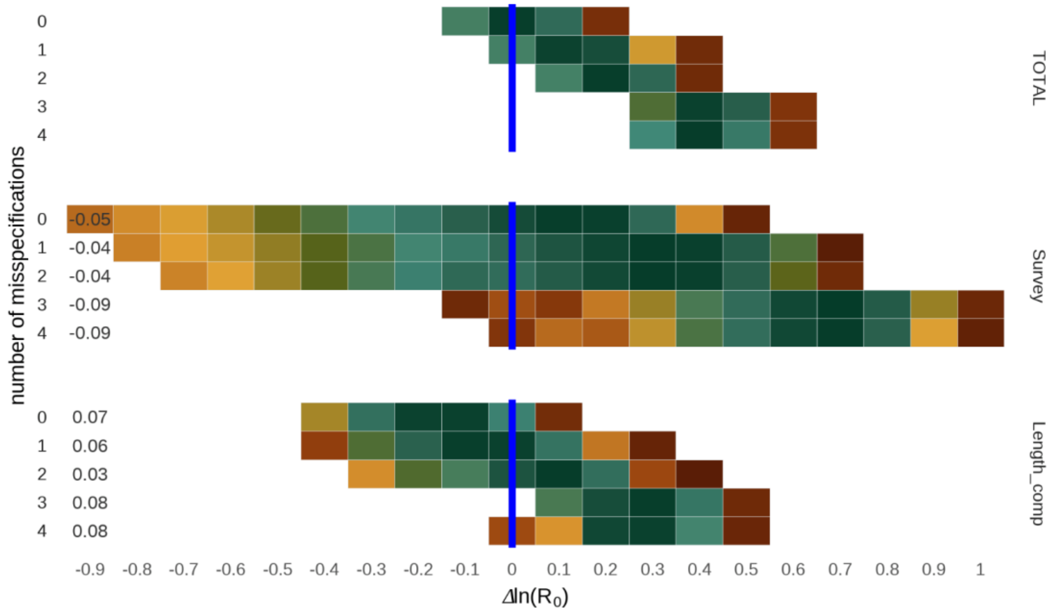


Figure 4. Likelihood profiles for log(R0) shown for a subset of estimation models with zero through four misspecifications. Each panel corresponds to either the total likelihood (top), survey likelihood (middle) or length composition (bottom). The x-axis has been recentered to the corresponding MLE from the correct (un-mis-specified) estimation model (vertical blue line); profiles have been filtered to only display model runs with changes in the negative log-likelihood less than 5 units. Green tiles indicate models closer to the minimum negative log-likelihood; red values are more disparate. The numbers on the left-hand side of the survey and length composition panels are the calculated psi statistic for each data component for each estimation model.

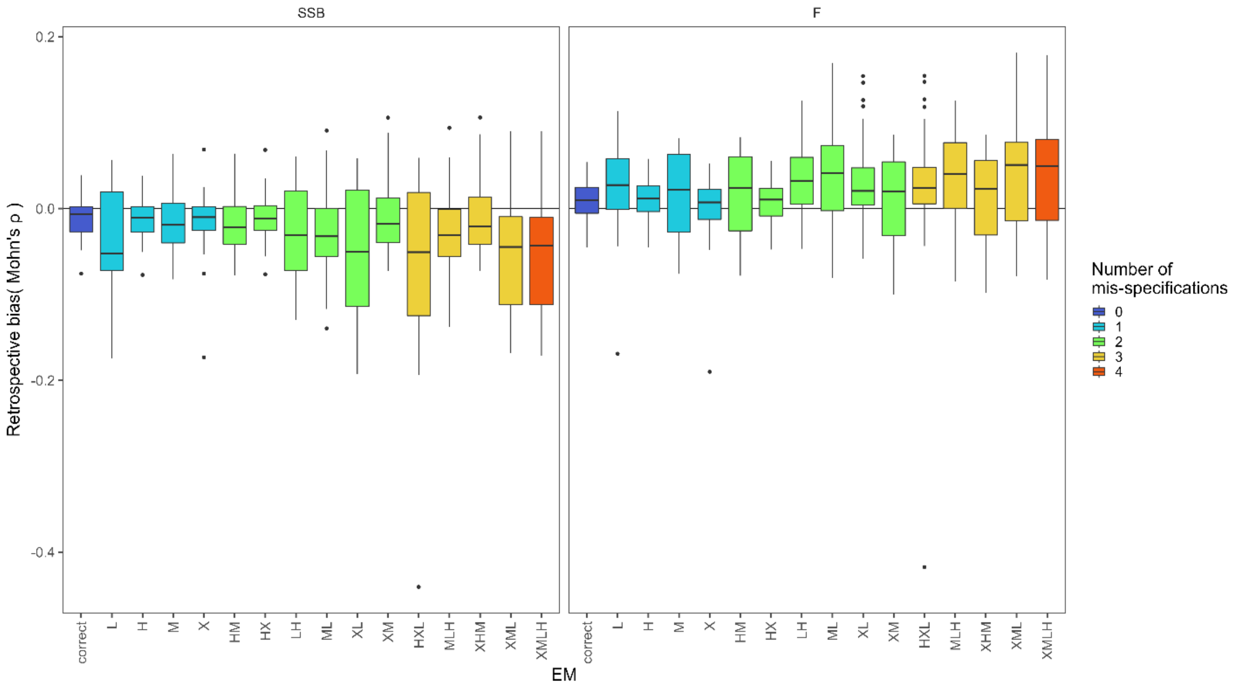


Figure 5. Boxplots of Mohn's rho from 5-year retrospectives in SSB (left) and F (right). The x axis labels indicate the parameters which are mis-specified. The colors indicate the total number of mis-specifications in a given model.

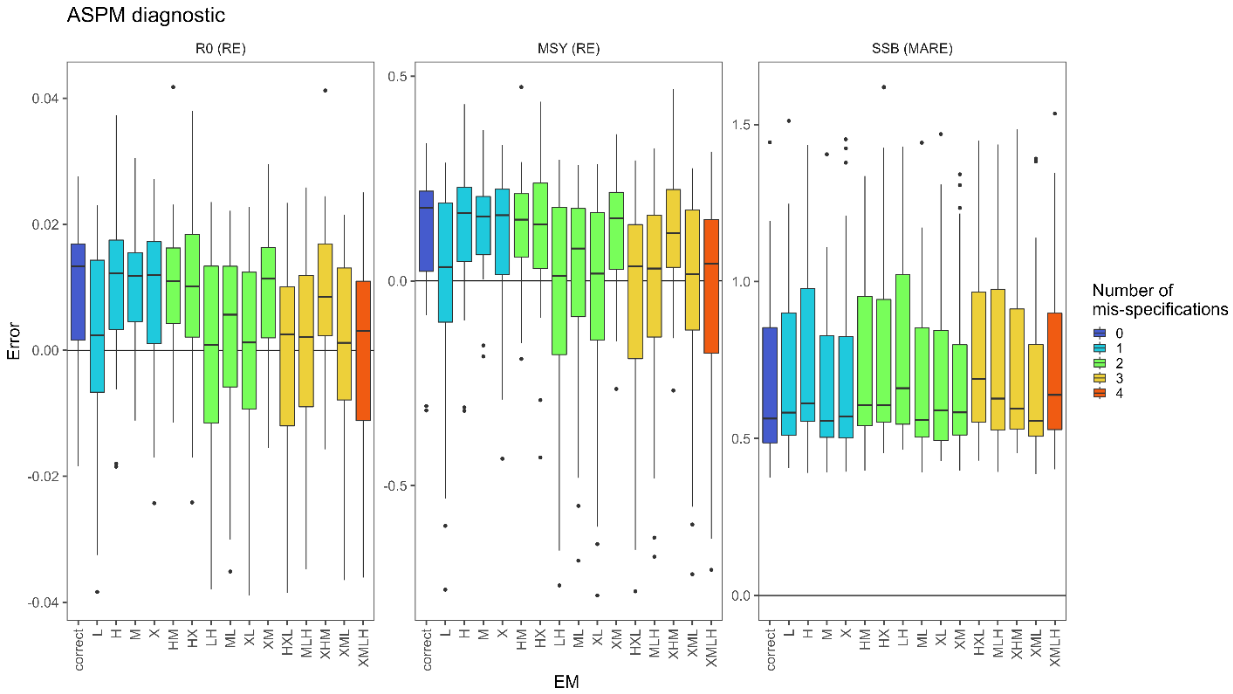


Figure 6. Boxplots of relative error in R0 (left), MSY (middle) and mean absolute relative error in SSB between the ASPM and full models. The x axis labels indicate the parameters which are mis-specified. The colors indicate the total number of mis-specifications in a given model.

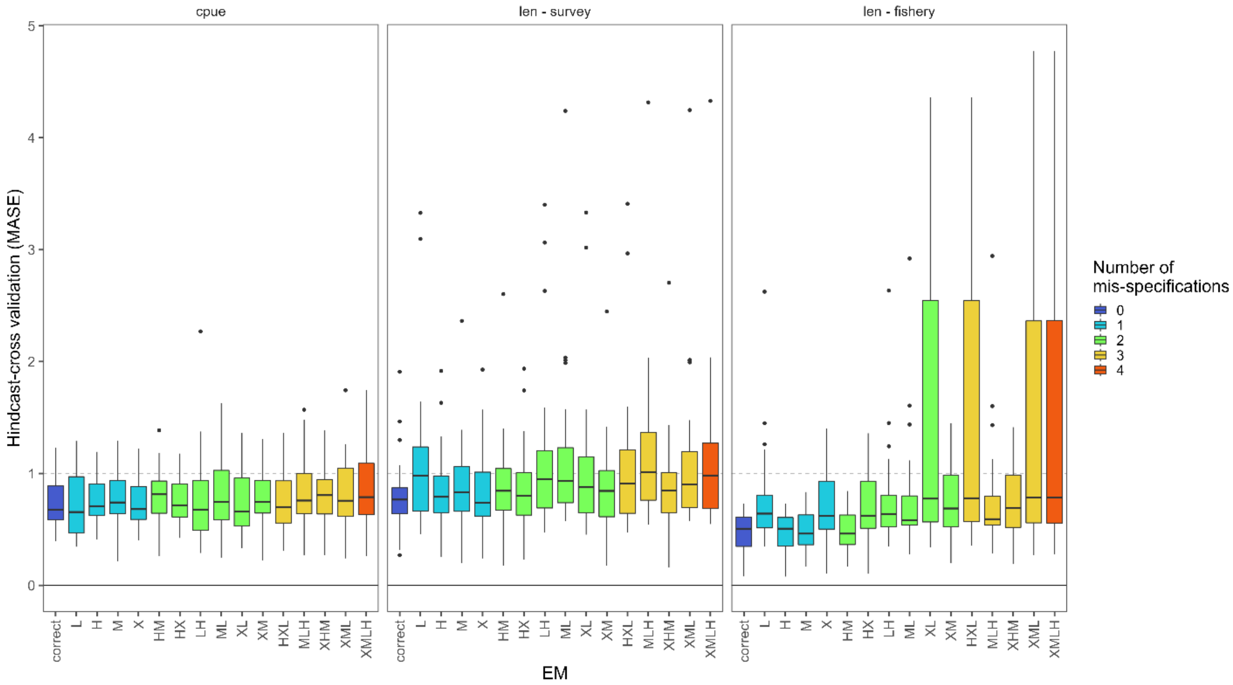


Figure 7. Boxplots of hindcast cross-validation MASE for the Survey CPUE (left), Survey length composition data (middle) and the fishery length composition data (right). The x axis labels indicate the parameters which are mis-specified. The colors indicate the total number of mis-specifications in a given model.

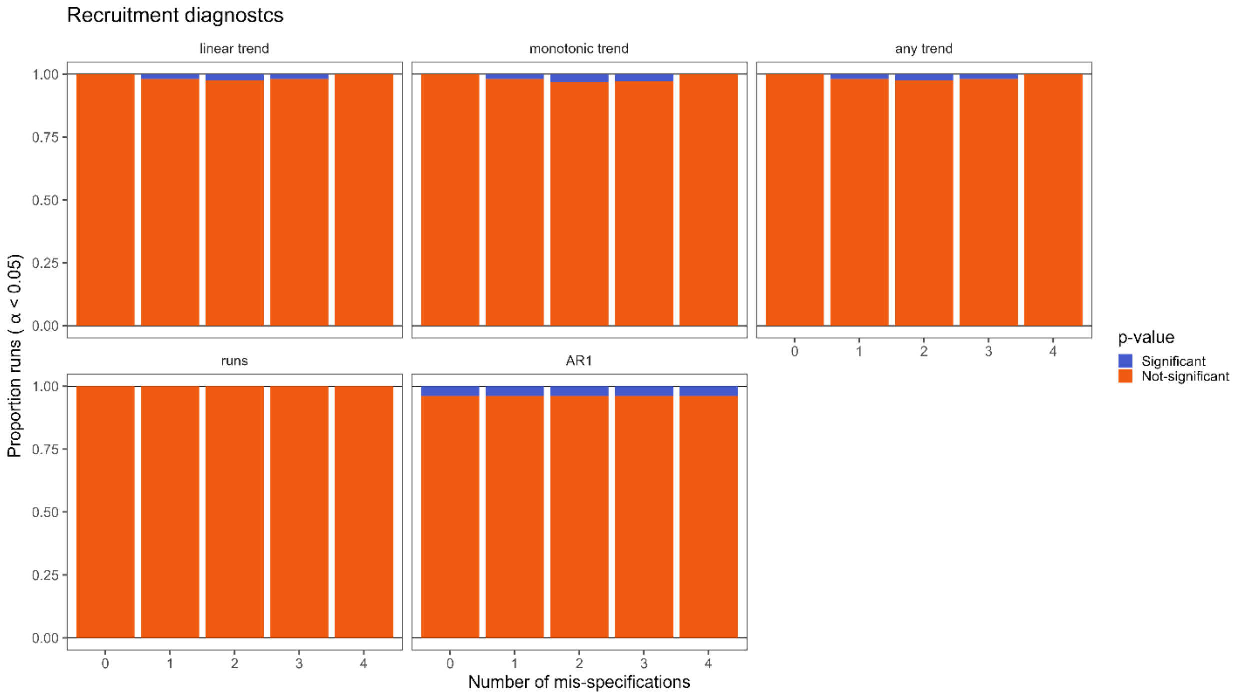


Figure 8. Proportion of runs by number of mis-specifications with significant results for tests of linear trend in the recruitment deviates (top-left), monotonic trend in the recruitment deviates (top-center), any trend in the recruitment deviates (top-right), non-random runs sequences in the recruitment deviates (bottom-left), or first order temporal autocorrelation in the recruitment deviates (bottom-center).

Supplementary material

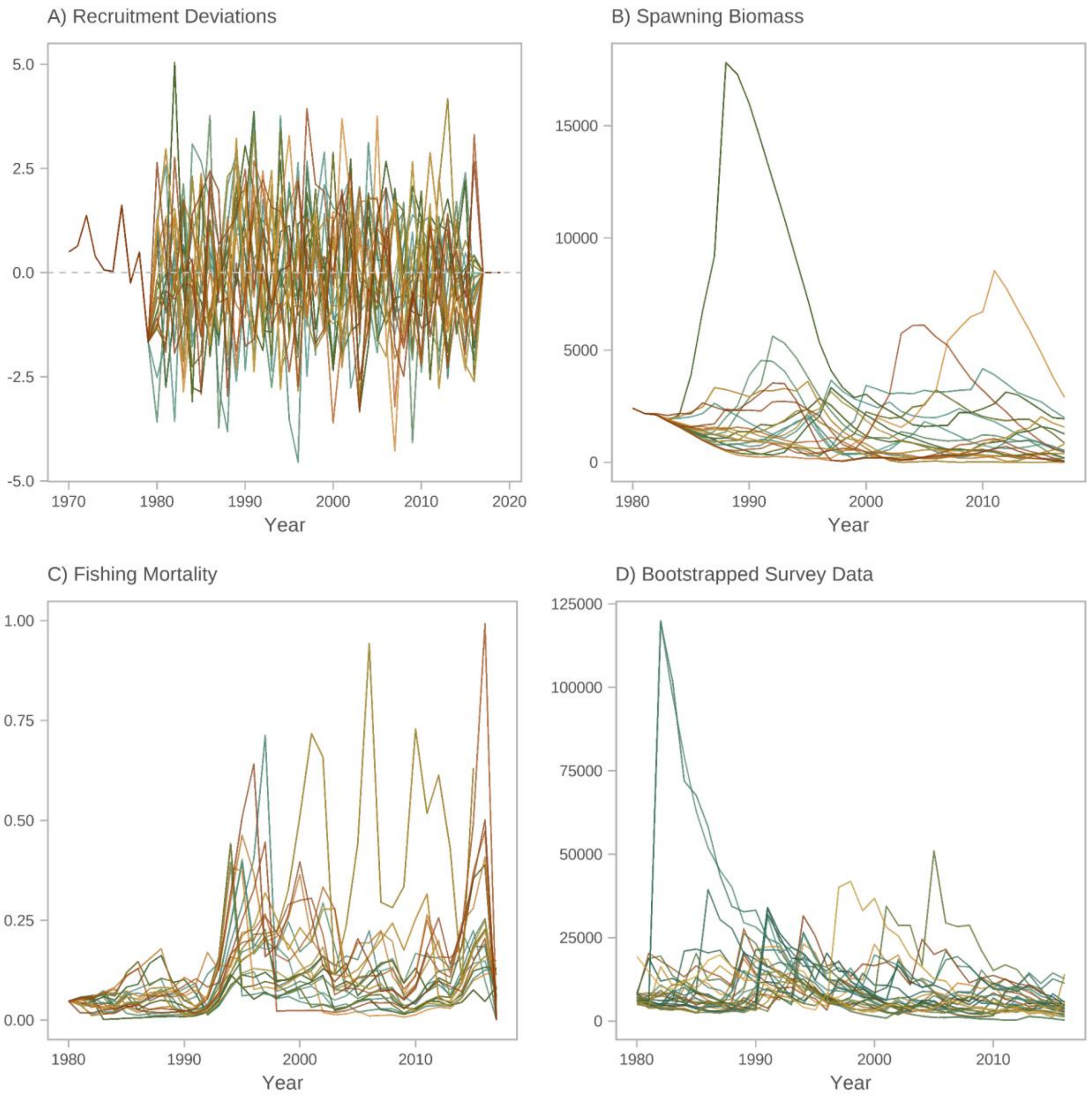


Figure S1. A) Log-scale recruitment deviations, B) Spawning stock biomass (in 1000 mt), C) apical fishing mortality, and D) indices of relative abundance (in 1000 mt) from 27 replicates of the operating model.

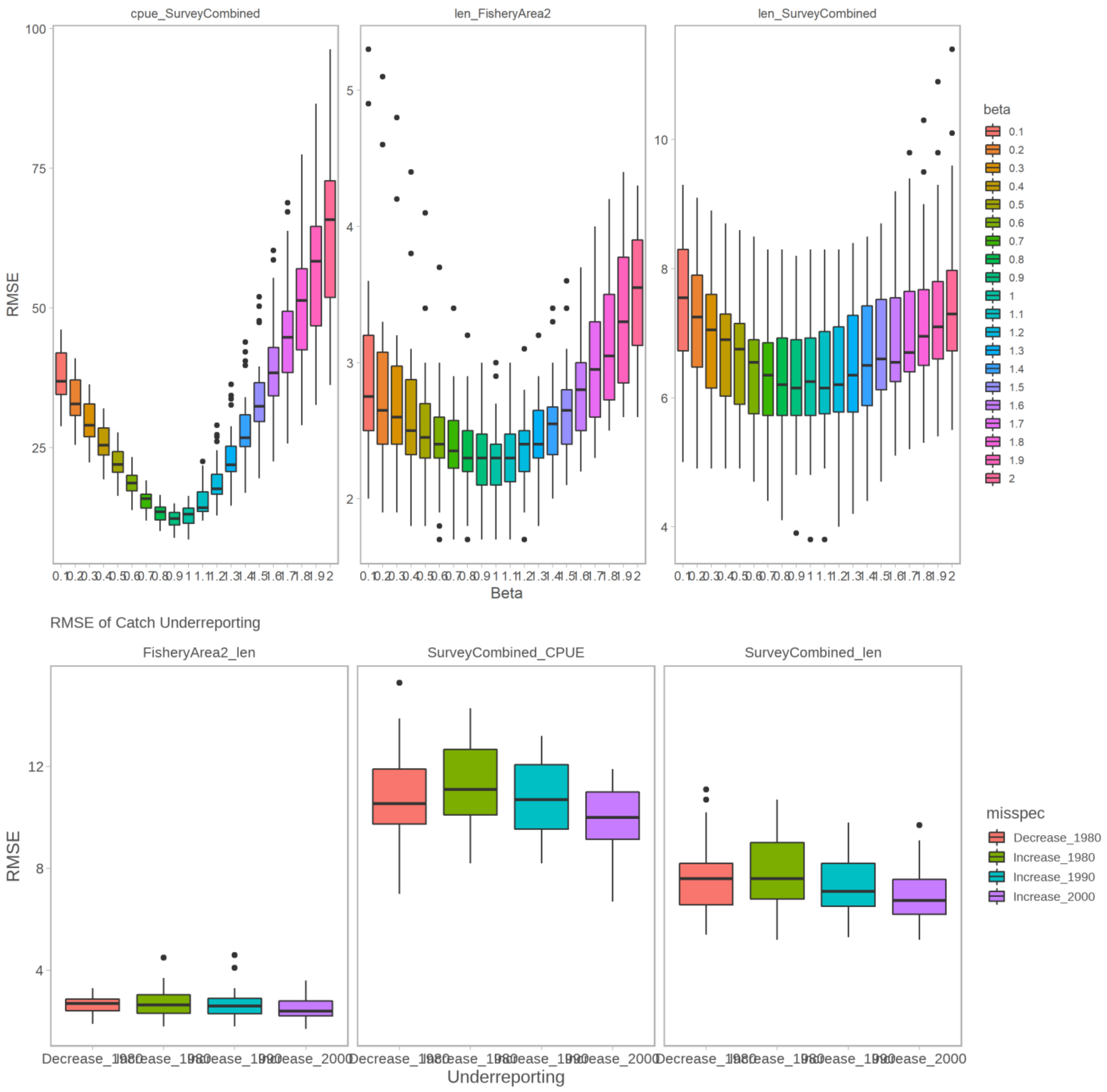


Figure S2. RMSE of survey and length composition data from data misspecifications hyperstability/depletion (top) and catch underreporting (bottom).

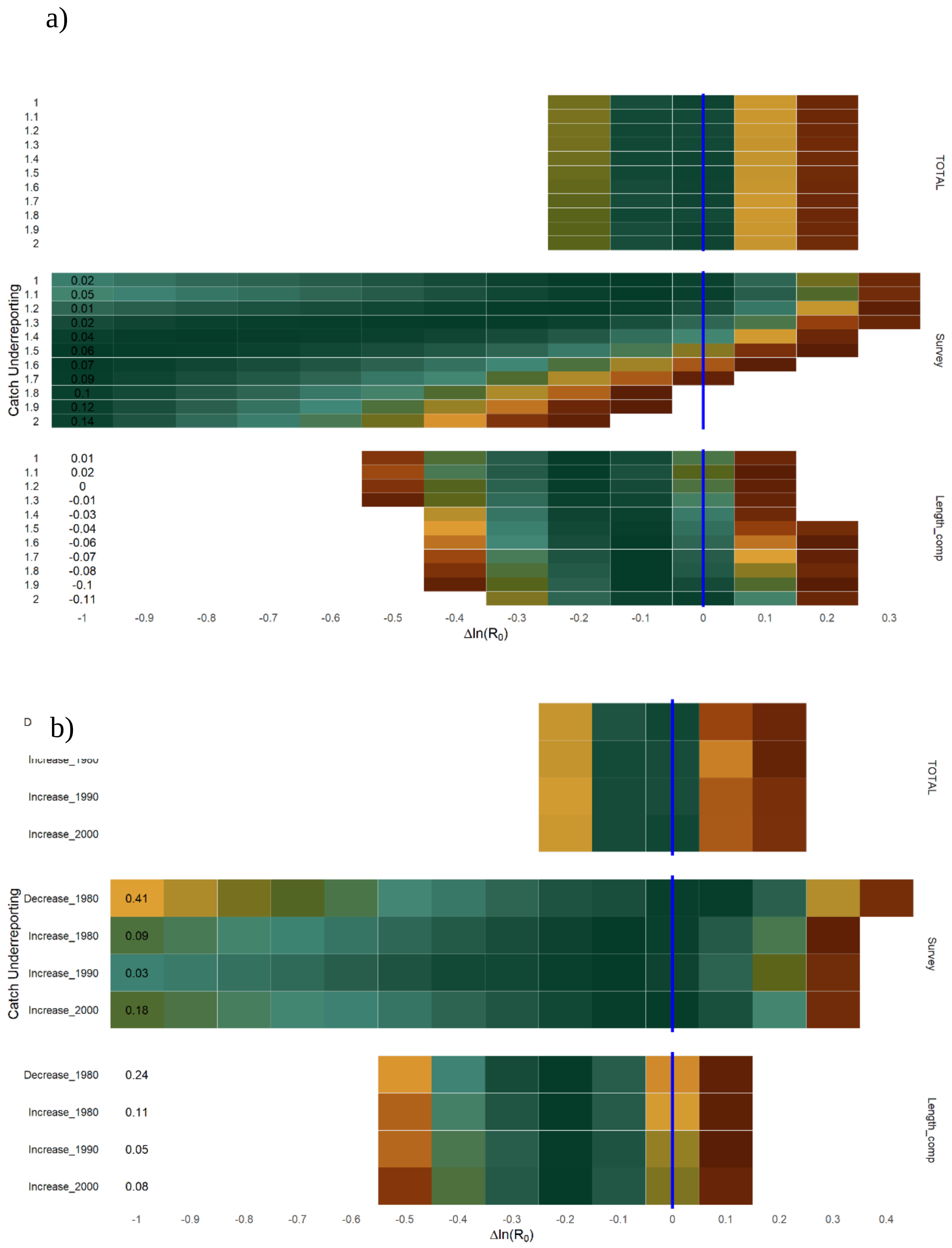


Figure S3. Likelihood profiles for log(R_0) shown for a subset of estimation models with hyperdepletion (a) or catch underreporting (b). Each panel corresponds to either the total likelihood (top), survey likelihood (middle) or length composition (bottom). The x-axis has been recentered to the corresponding MLE from the model with correct (un-mis-specified) data (vertical blue line); profiles have been filtered to only display model runs with changes in the negative log-likelihood less than 5 units. Green tiles indicate models closer to the minimum negative log-likelihood; red values are more disparate. The numbers on the left-hand side of the survey and length composition panels are the calculated psi statistic for each data component for each estimation model.

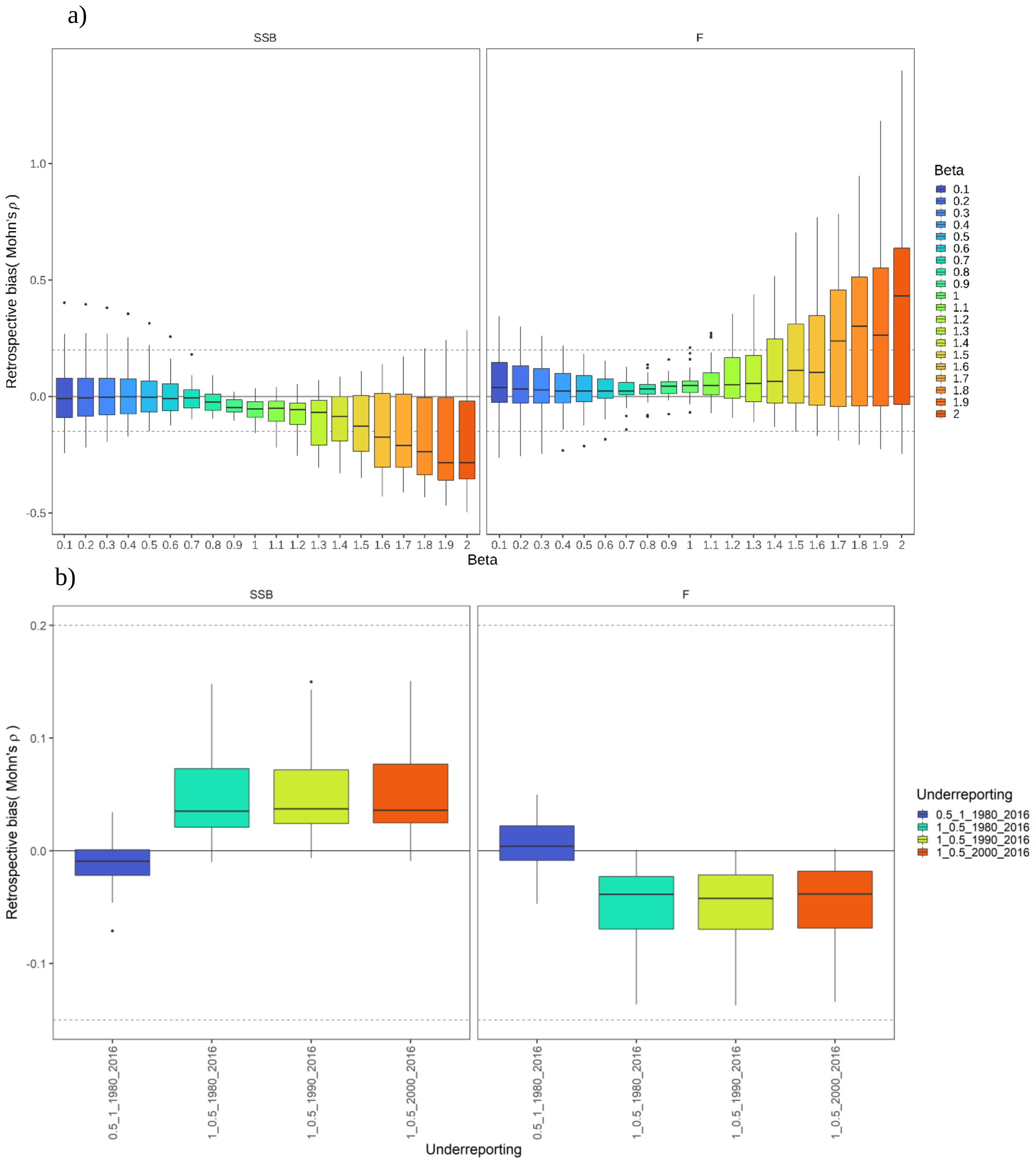


Figure S4. Boxplots of Mohn's rho from 5-year retrospectives in SSB (left) and F (right). The x axis labels indicate the level of hyperstability or hyperdepletion (a) or the underreporting scenario (b).

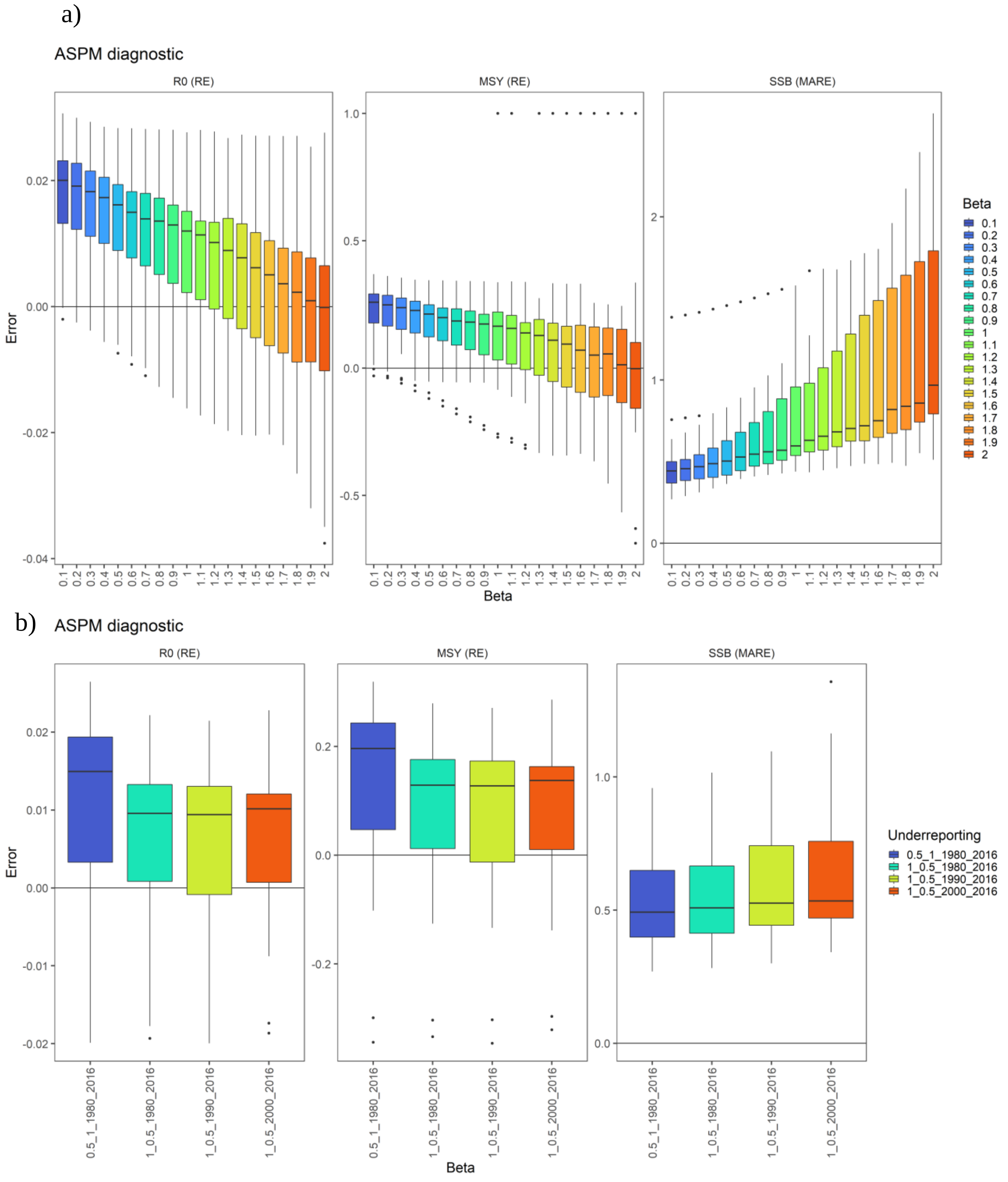


Figure S5. Boxplots of relative error in R0 (left), MSY (middle) and mean absolute relative error in SSB between the ASPM and full models. The x axis labels indicate the level of hyperstability/depletion (a) or catch underreporting scenario (b).

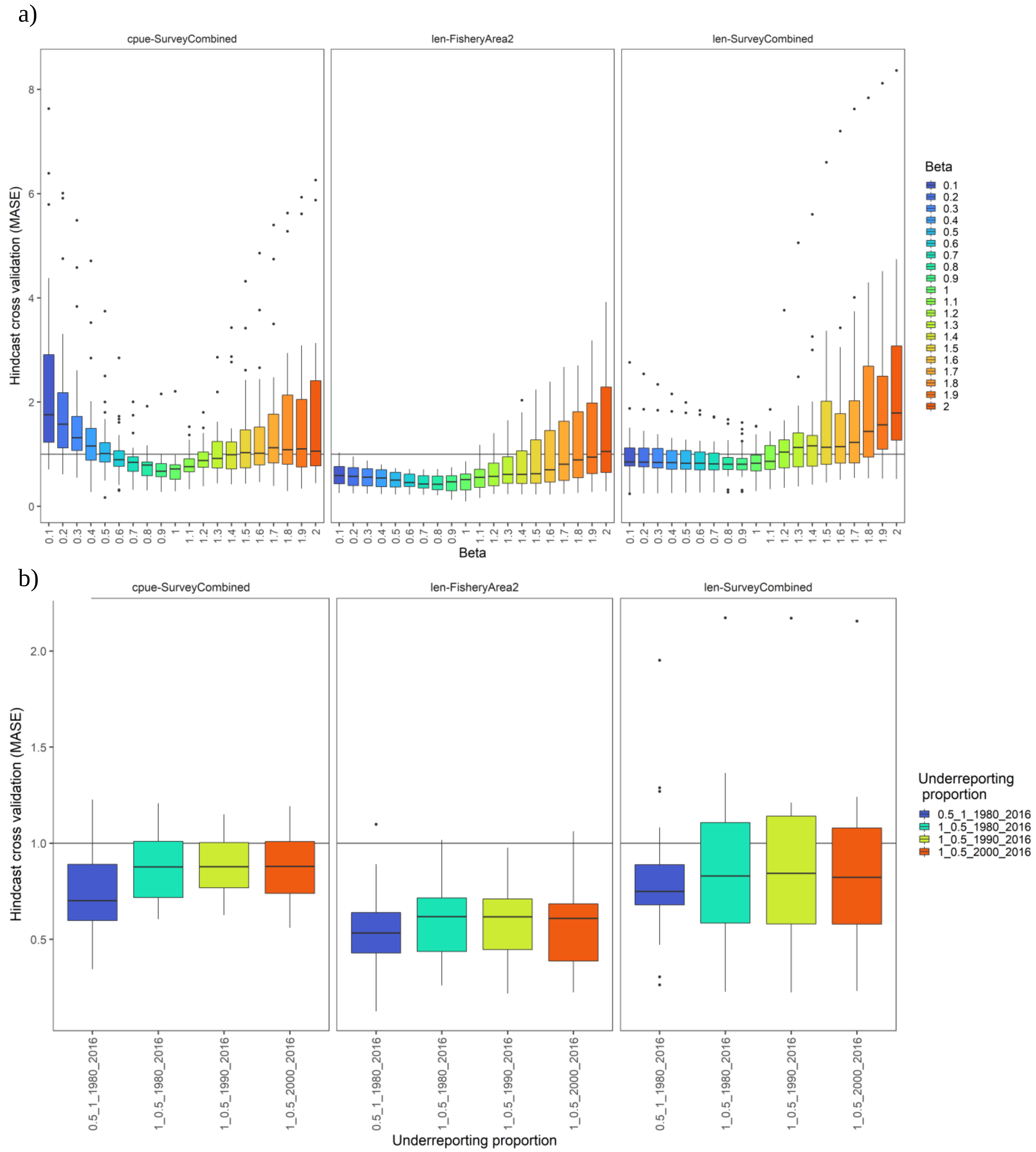
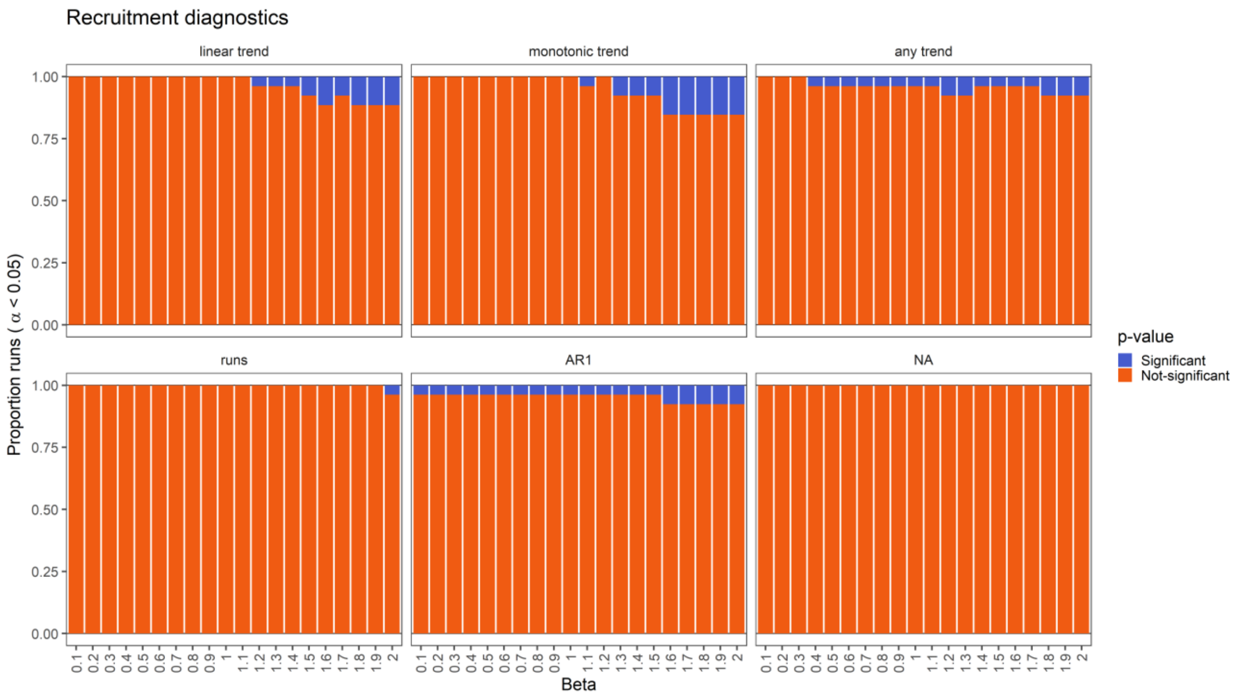


Figure S6. Boxplots of hindcast cross-validation MASE for the Survey CPUE (left), Survey length composition data (middle) and the fishery length composition data (right). The x axis labels indicate the level of hyperstability/depletion (a) or catch underreporting scenario (b).

a)



b)

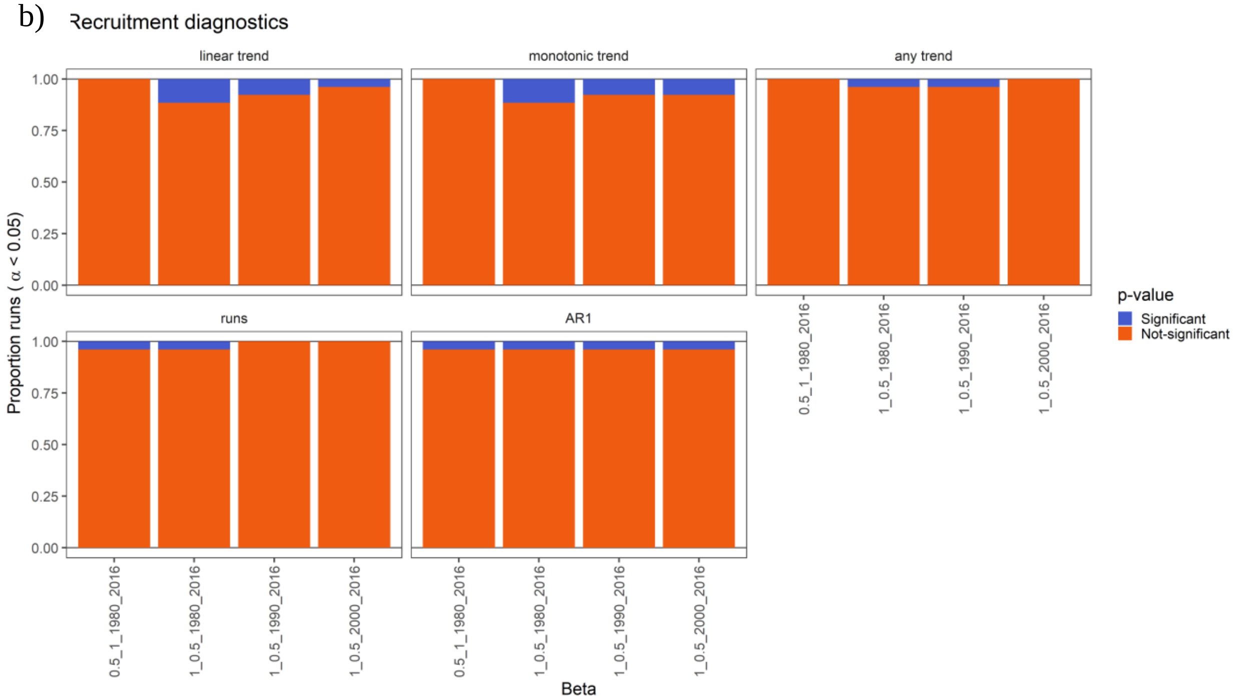


Figure S7. Proportion of runs by number of mis-specifications with significant results for tests of linear trend in the recruitment deviates (top-left), monotonic trend in the recruitment deviates (top-center), any trend in the recruitment deviates (top-right), non-random runs sequences in the recruitment deviates (bottom-left), or first order temporal autocorrelation in the recruitment deviates (bottom-center) by hyperstablity/depletion (a) or catch underreporting (b) misspecifications.