A state-space assessment of American plaice using the Woods Hole Assessment Model (WHAM)

Amanda Hart, Lisa Kerr and Tim Miller

4.1 Introduction

The American plaice stock assessment has been conducted using a Virtual Population Analysis, VPA, since 1992 (NEFSC 1992). The model can be obtained from the NOAA Fisheries Toolbox (https://nmfs-fish-tools.github.io/VPA/). The current version of stock assessment model for American plaice was developed in GARM III (NEFSC 2008) and updated in subsequent operational stock assessments (NEFSC 2012, NEFSC 2015, NEFSC 2017, NEFSC 2019). There was a desire in this research track stock assessment process to switch from the VPA to either a statistical catch-at-age framework (ASAP) or the state-space model, WHAM.

4.2 WHAM Overview

The Woods Hole Assessment Model (WHAM; Stock and Miller 2021; https://github.com/timjmiller/wham), is a state-space age-structured stock assessment model that can be configured in a similar manner to the Age Structure Assessment Program (ASAP; Legault and Restrepo 1999) with fits to aggregated catch, index, and age composition data. WHAM also provides several alternative age composition models, and can include process errors and environmental covariates.

Development of a WHAM model for American plaice began with bridge and scaling runs to explore the implications of switching from a VPA model to WHAM with updated model assumptions. Runs with alternative configurations of selectivity, age composition likelihoods, and abundance at age were explored with updated data for the fleet and the survey indices previously included in the VPA (Northeast Fisheries Science Center [NEFSC] spring and fall bottom trawl, 1980-2019). These base runs were expanded to explore an alternative catch model that was fit to an extended catch time series beginning in 1960, and the inclusion of Massachusetts Division of Marine Fisheries (MADMF) and Maine New Hampshire (ME-NH) inshore trawl surveys, a landings per unit effort (LPUE) index, Vector-Autoregressive Spatio-Temporal (VAST) model-based indices, and a split in the NEFSC indices between Bigelow and Albatross vessel years. Finally, runs built upon the above runs and explored environmental covariate links to recruitment and catchability. Model results and diagnostics for all runs (including exploratory runs not discussed in detail here) are available on GitHub (https://github.com/ahart1/PlaiceWG2021).

Candidate models are discussed briefly in the above context and in detail in the Section 4.12. Reference points and short-term projections are also provided for each candidate model.

4.3 Model diagnostics

Five diagnostics were used to compare model fit and performance to identify the candidate WHAM models and are briefly described here. All models were assessed for first and second order convergence. Models passed first order convergence criteria when their final gradient was smaller than 1e--10 and second order convergence if the hessian was invertible. Akaike's Information Criteria (AIC; Akaike 1974) was used to compare models with the same likelihood structure (e.g., models fit to the same data). Smaller AIC scores indicated model improvement and scores within +/-2 of each other were considered equivalent, in which case the more parsimonious of the two models were selected. Mohn's rho values (Mohn 1999) were used to identify retrospective patterns in recruitment (R), spawning stock biomass (SSB) and fully recruited fishing mortality (Fbar), where smaller absolute values are preferred. Mohn's rho values were comparable across all WHAM runs that met convergence criteria, while AIC scores were only compared for runs with the same likelihood structure and are thus color-coded to show comparable runs in Table 4.1. Residuals were used to assess fit to data, and residual patterns helped identify possible sources of model misspecification to be addressed in subsequent runs. Early runs used observed minus predicted residuals to compare performance, but all model runs after 25 were compared using one-step-ahead (OSA) residuals that are more informative for state-space models (Thygesen et al. 2017). OSA residuals should be uncorrelated and normally distributed for models that appropriately describe the system and thus provide an additional diagnostic for assessing model appropriateness beyond looking for any residual patterns. Mean absolute scaled error (MASE) scores were calculated for possible candidate WHAM models to compare differences in model and index prediction skill (Carvalho et al. 2021; Kell et al. 2021). Values below 1 indicated models with predictive skill that is better than a naive approach, and a MASE score of 0.5 has twice the predictive skill as a naive approach. Finally, simulation self-tests were performed for the three candidate models. Each model was used to generate 100 datasets with parameters fixed at their estimated values. Simulations then refit the model to these generated datasets to evaluate relative error in F, SSB, R, and catch estimates and model convergence rates.

4.4 Bridge Runs

Two bridge runs were conducted to configure WHAM as closely as possible to the 2019 VPA model (NEFSC 2008, 2019). Data from the VPA (1980-2018) was imported into an ASAP data file for use in both the ASAP (Alade or Cadrin reference ?) and WHAM bridge runs. The first bridge run used identical input data and model configuration as the

ASAP bridge run and includes age-disaggregated indices for all ages in the NEFSC spring and fall bottom trawl surveys. The second bridge run used the same data but implemented a single age-aggregated index for both the NEFSC spring and fall surveys.

4.5 Scaling Runs

Three additional WHAM runs were conducted to explore the consequences of updating model data and changing natural mortality (M) and maturity assumptions in this research track. An additional year of catch (1980-2019), NEFSC spring (1980-2019) and fall survey (1981-2019), and weight-at-age (1980-2019) data were added to existing time series from the VPA. In addition, fleet discards were estimated rather than imputed and the age classes for NEFSC indices were expanded to include an 11+ group rather than the 9+ used in the VPA. Natural mortality was assumed constant across ages and throughout the time series, but was increased from 0.2 assumed in the previous VPA assessment to 0.3 in this research track to reflect revised mortality expectations. Maturity-at-age was also revised from annual estimates to a constant maturity schedule throughout the time series.

One model (run 9) was conducted with all updated data, M, and maturity assumptions. This was compared to a second model (run 10) with updated data and maturity assumptions but reverted to the VPA M assumption (0.2) and a third model (run 11) with updated data but reverted to VPA M (0.2) and the annually varying maturity expectations used in the VPA. Reverting to the VPA M scaled SSB, F, and R estimates downward compared to the run with all data updated, but changes in the maturity assumption did not scale model estimates as dramatically and this third model performed similarly to the second run where only M was reverted to the VPA setting (Figure 4.1).

4.6 Alternative Selectivity Assumption Runs

Selectivity runs began by exploring selectivity models for the fleet and indices. Fleet selectivity was assumed to be logistic for all runs, but both age-specific and logistic selectivity was explored for the NEFSC spring and fall indices. For runs that assumed age-specific index selectivity, NEFSC spring ages 4 and 5 and NEFSC fall age 4 selectivity were fixed at 1 based on a preliminary run that freely estimated selectivity for all ages (run 12). Two fleet selectivity blocks (1980-1999 and 2000-2019) were implemented in run 16 to account for differences between historic and contemporary fishing patterns associated with a series of mesh size increases in the late 1990s and early 2000s and to break the time series roughly in half to avoid assuming constant selectivity for the entire time series. These selectivity blocks improved the fit to catch

data and lowered the AIC score. Age-specific selectivity estimates were dome-shaped except for the NEFSC fall survey which exhibited a higher estimated selectivity for the 11+ group than for age 10 fish. Run 16A explored the alternative use of logistic selectivity for both indices, but this resulted in a higher AIC and likelihood contributions with few changes to model expectations (Figure 4.2). Time-varying selectivity was also explored by implementing independent and identically distributed (iid) random effects for the fleet and both indices (run 23), resulting in a better fit to the data and smaller age composition residuals. Limiting the selectivity random effect to only the fleet (run 23A) resulted in a slight reduction in the Mohn's rho values for SSB and Fbar, but a higher AIC with little difference in the fit to catch data. Indices were not fit as well and had higher variability in selectivity for ages 2-3 when random effects for index selectivity were not included. Assuming an autocorrelated (AR1) process for selectivity random effects (run 24) did not improve model fit, with larger fleet age composition residuals and retrospective patterns.

4.7 Abundance-at-age Random Effects Runs

Model runs treating recruitment deviations as independent random effects built on run 23 with selectivity deviations treated as iid random effects. Assuming iid random effects for recruitment (run 25) reduced the CVs of recruitment estimates, particularly towards the end of the time series, but otherwise performed similarly to runs without recruitment random effects. Assuming recruitment deviations are autocorrelated by year (AR1 random effects, run 26) resulted in a slightly smaller AIC and Mohn's rho for SSB and Fbar but otherwise very similar performance (see Table 4.1). Moving to a full state-space model (run 27) with random effects for all ages resulted in a much smaller AIC and Mohn's rho for R, SSB, and Fbar than for runs with only a recruitment random effect, however CVs around recruitment estimates were much higher. General trends in F and selectivity were similar to runs 25-26, but the scale differed in some years and selectivity for the NEFSC fall index was less variable than in previous runs. Catch residuals for this run were smaller in magnitude for some years and slightly more evenly distributed around zero.

One-step-ahead (OSA) residuals are more informative than response residuals (observed minus predicted values) for state-space models because they are uncorrelated and should be normally distributed if the model appropriately describes the system (Thygesen et al. 2017). For this reason, OSA residuals for fit to catch and index data were calculated for all runs, and age composition OSA residuals were also calculated for run 25 and runs 27-50A (available through the development version of WHAM on GitHub). Run 27 had improved fits to age composition data (smaller OSA residuals), but older fish tended to have more negative residuals than younger fish.

4.8 Extended Catch Time Series Run

Catch data for American plaice is available as early as 1960, but no age composition data is available prior to 1980 so a run was conducted to examine the impact of including the extended catch time series and several additional years of NEFSC spring (1968-2019 with age composition available beginning in 1980) and fall (1963-2019 with age composition available beginning in 1980) index data (run 28). The full state-space model did not converge, but a model with recruitment iid random effects (as in run 25) did converge and exhibited similar model performance for years in which both models had data (1980-2019). Residuals for fits to catch and index data prior to 1980 tended to be positive, but after 1980 had similar trends and magnitudes to those from run 25. Given similar fits to data and estimates of SSB, F, and recruitment, reference points for runs with and without the extended time series (runs 28 and 25 respectively) were used to identify potential differences in management performance between these two runs that otherwise have very similar model diagnostics. F40% (the fishing mortality that is expected to maintain 40% of the maximum spawning potential under no fishing) was similar across both runs, but run 28 had much higher expectations for SSB at F40% and Yield at F40% and wider confidence intervals around these expectations with run 25 SSB and yield expectations falling near the lower confidence interval for run 28 (Figure 4.3).

4.9 Alternative Index of Abundance Runs

4.9.1 Splitting the Albatross and Bigelow time series

A full state-space model (run 29) with iid selectivity random effects was fit to examine the impact of splitting the NEFSC spring and fall indices into separate Albatross (1980-2008) and Bigelow (2009-2019) indices. The fit to the fleet age composition data was slightly worse (i.e., larger likelihood contribution) and fleet OSA residuals for fit to aggregate catch data were less normally distributed than the run with the combined NEFSC indices (run 27). Age composition residuals calculated as observed minus predicted values showed slight improvement for age 1, particularly in the NEFSC fall index, but OSA residuals for fit to aggregate indices indicated similar residual patterns and magnitudes for fits to combined (run 27) and split (run 29) surveys. OSA residuals for fit to age composition data were fairly normally distributed for the fleet and aggregated spring and fall indices in run 27, with some differences in run 29 due to the split between Albatross and Bigelow years. In run 29 these residuals were reasonably normally distributed for the fleet and all indices, but indices had age-specific residual patterns with generally positive residuals for ages 5-11+ particularly later in the Albatross time series and a better mix of positive and negative for younger ages. Bigelow fall residuals were also a bit less normally distributed than for other age

composition data. AIC for run 29 was smaller as were Mohn's rho estimates for SSB, Fbar, and recruitment so model selection favors this model over run 27. Estimated SSB was higher, F was slightly lower, and recruitment was slightly higher than in run 27 with larger differences toward the end of the timeseries (Figure 4.4).

Run 29 leveraged age composition in numbers at age, and Albatross calibrated data during both Albatross and Bigelow years (i.e. split the combined NEFSC spring and fall indices without removing the calibration to Albatross units from 2009-2019). Three further runs (29A-29C) explored the consequences of using biomass age composition data and uncalibrated Bigelow time series (i.e. no calibration to Albatross units was used) from 2009-2019. Run 29A implemented a full state-space model with iid selectivity random effects for the fleet and Albatross spring index, but fit to biomass age composition data and used data in Bigelow units from 2009-2019 . The selectivity pattern for the Bigelow fall index was more dome-shaped than in run 29, and did not have an increase in estimated selectivity from age 10 to 11+ as was previously estimated. Selectivity random effects for the Albatross spring index were less variable than in run 29, indicating that they could potentially be removed (explored in run 29B). Catchability estimates for Bigelow spring and fall and the Albatross fall indices were more similar than in run 29, although the Albatross spring catchability estimate was much lower than for the other three indices. Catch residuals were larger in magnitude for run 29A but were a more even mix of positive and negative values than in run 29. The fleet and spring indices had similar OSA age composition residual patterns but had larger maximum residuals with most differences for ages 1-3, while the fall indices had smaller maximum residuals. OSA residuals for fit to the fleet and Albatross fall index were more normally distributed (i.e. a better fit) than in run 29, with other indices showing a similar distribution in OSA residuals.

Run 29B implemented the same model as in run 29A but excluded selectivity random effects for the survey indices. Model diagnostics were very similar to run 29A. Mohn's rho values for recruitment and Rbar were slightly smaller (improved) while the value for SSB was slightly larger (worse performance). However, the delta AIC values were < 2 so these models should be considered equivalent, and the simpler model (run 29B) with selectivity random effects only for the fleet would be the proposed model for further development. Run 29B-1 implemented the same model but made slightly different assumptions for index effective sample size based on those used in ASAP run 51a. This adjustment had very little impact on model estimates or fit.

Run 29C implemented the same model as in run 29A but fit to biomass age composition data and used Albatross units for both the Albatross and Bigelow years to identify whether the improvements in run 29A were attributed to the switch to biomass age

composition data, or to the switch from Albatross to Bigelow units for 2009-2019. Selectivity estimates for age 11+ in both fall indices were higher than age 10 so the improvement in the Bigelow fall index in run 29A and 29B appears to be attributable to the switch to Bigelow units. The selectivity random effects for the Albatross spring index was smaller than in run 29, as was seen in run 29A and suggesting that removing this random effect for model fit is appropriate (i.e. run 29B is also an improvement over this run). Catch residuals were more evenly distributed around zero early in the time series as was also seen in runs 29A and 29B but not in run 29. This suggests that this improvement is due to the switch to biomass units rather than the switch to Bigelow units. OSA age composition residuals for the fleet and spring indices had similar patterns but larger maximum values compared to run 29. The fall indices followed similar patterns but had slightly smaller maximum values. The patterns and magnitude of these residuals were very similar to those in run 29B.

4.9.2 Incorporating State Indices of Abundance

Models were fit to combinations of NEFSC, inshore state trawl (MADMF and ME-NH), LPUE, and VAST indices built upon run 27 to explore other indices of abundance available for American plaice. A full state-space model was fit to NEFSC and MADMF spring and fall indices (run 30) but did not include selectivity random effects for the indices because the model failed to converge (i.e., hessian was not invertible) when they were included. The exclusion of selectivity random effects was expected to have the biggest impact for ages 1-3 of the NEFSC spring index, and these ages tended to have the largest one-step-ahead (OSA) age composition residuals. The fit to fleet and NEFSC spring age composition was worse, with only minor improvement in fit (i.e., smaller maximum residuals) to the NEFSC fall index. Residuals for fit to the NEFSC indices were consistently positive at the end of the time series while residuals for fit to the MADMF indices were consistently negative. This pattern was also observed for runs that included ME-NH indices. AIC was not comparable to run 27, but the Mohn's rho value for F was smaller than run 27, while values for SSB and Fbar were slightly larger.

A full state-space model with selectivity random effects converged when fit to both NEFSC and ME-NH spring and fall indices (run 31). The ME-NH inshore trawl survey fully selected much younger ages (ages 2 and 1 for the spring and fall respectively) compared to the NEFSC indices (ages 4-5 in the spring and age 4 in the fall), but had little impact on residual patterns for ages 1-2 in run 31. Fleet age composition OSA residuals were slightly smaller in run 31 than in run 27 but followed a similar pattern, while the NEFSC spring and fall indices had larger residuals, particularly for ages 1-3. Fits to the ME-NH age composition followed a similar pattern with large residuals observed in many years for ages 1-3 and generally smaller residuals for older ages.

Mohn's rho values for R and SSB were smaller than in run 27, but the value for Fbar was slightly larger.

4.9.3 Incorporating Fishery Dependent Indices of Abundance (LPUE)

A model with recruitment random effects converged when fit to NEFSC and LPUE indices (run 32). The selectivity pattern for the LPUE index was specified to mirror the fleet's estimated selectivity-at-age. Fit to catch data was poorer than the run (run 25) that only fit to NEFSC indices (i.e. larger extreme catch residuals, less normally distributed OSA residuals), but the fit to NEFSC indices was similar. OSA residuals for fit to age composition data were similar between run 25 and this run for both the fleet and the NEFSC indices. Mohn's rho values for SSB and Fbar were smaller in run 32 than in run 25, but the Mohn's rho value for R was larger. An additional sensitivity run (32A) was conducted to fit to the same indices but implemented a catchability random effect for the LPUE index. This resulted in slightly more normally distributed OSA residuals for fit to the fleet and aggregate indices, and much more normally distributed OSA residuals for the LPUE index. Despite these improvements the inner-quartile range of MASE scores for this run was much broader than for runs 27 and 29-29H and at a prediction horizon of 1 year the mean is greater than 1 (on average less accurate than a mean approach). The MASE score describing the prediction skill of the LPUE index in run 32A also had a mean above 1 suggesting that this index does not improve the prediction skill of the model. Mohn's rho values were not calculated for run 32A.

4.9.4 Incorporating Multiple State and Fishery Dependent Indices of Abundance

Three runs explored the consequences of including multiple state indices in the assessment model. Run 33 fit a full state-space model to NEFSC, MADMF, and ME-NH spring and fall indices but did not include selectivity random effects. Fit to the fleet and NEFSC spring index were similar to run 30 that fit only the NEFSC and MADMF indices, but the fit to the NEFSC fall index was worse, particularly for ages 1-2. Fit to the MADMF spring index was worse with very large OSA residuals for ages 1-2 in some years, and fit to the MADMF fall index had slightly smaller maximum OSA residuals. The ME-NH indices had very large residuals for ages 1-3. Despite these differences, the magnitude and pattern of SSB and F estimates were very similar to run 30, with more variability in recruitment estimates between runs. Similarly, estimates of F40% are similar to run 30 but estimates of SSB and yield at F40% fell between those for run 30 (fit to NEFSC and MADMF) and run 31 (fit to NEFSC and ME-NH, Figure 4.5). Mohn's rho values for R, SSB, and Fbar in run 31 were all larger than for run 30.

Two runs (34A and 35A) explored models fit to a combination of NEFSC, state, and LPUE indices. Similarly to run 32 which fit to both the NEFSC and LPUE indices, the LPUE index selectivity was specified to mirror the fleet selectivity estimates for these

runs. Both runs had a poorer fit to catch data (i.e. larger catch residuals and less normally distributed aggregate catch OSA residuals) compared to runs that fit to only the NEFSC and state indices (runs 31 and 30 compared, respectively, to 34A and 35A). OSA residuals for fit to the aggregate LPUE index were not very normally distributed for either run, indicating that these models are not entirely appropriate for this data. Age composition OSA residuals for the NEFSC spring and fall indices were both slightly more normally distributed than in runs 31 and 30, indicating a slight improvement in fit to these indices. Mohn's rho values for SSB and Fbar were smaller in run 34A than in run 31 but larger for R. In contrast, all Mohn's rho values for run 35A were smaller than in run 30. These runs converged when a recruitment random effect was specified but did not converge when random effects for all numbers-at-age were implemented, thus they were not considered further in this analysis.

4.9.5 Incorporating Model-Based Indices of Abundance (VAST)

Run 37E replaced design-based indices with fits to several iterations of model-based Vector-Autoregressive Spatio-Temporal (VAST) indices. Earlier iterations (37-37D) were developed using preliminary VAST index data and are thus not discussed in detail here. VAST analysis was used to generate both spring and fall indices based on raw NEFSC, MADMF, and MENH trawl data in uncalibrated units (i.e. no survey units calibrated to Albatross units), and covered numbers-at-age for 1-11+. Run 37E fit to these updated VAST indices, and assumed logistic selectivity for the fleet and both spring and fall indices based on a preliminary run that freely estimated selectivity-at-age. OSA residuals for fit to the aggregate catch were less normally distributed than in run 29B, but OSA residuals for fit to fleet age composition data were generally more normally distributed with the exception of some very large residuals that were attributed to age 10 and were generally large and positive. OSA residuals for fit to aggregate indices were more normally distributed in the fall than in the spring and OSA residuals for fit to age composition data showed a similar trend as the fleet, with age 10 residuals often very large and positive, but otherwise a fairly normal distribution. Observed-predicted residuals for the indices showed a better fit than prior VAST runs (e.g. 37B), indicating that poor fit in those runs could be a data issue. This run could not be directly compared to other runs via AIC due to differences in the input data, but Mohn's rho values for R, SSB, and F were larger than those in runs 29B and 29F.

4.10 Alternative Age Composition Runs

Two additional sets of runs explored alternative age composition models (logistic-normal, dirichlet-multinomial) that more explicitly weight indices (Fisch et al. 2021) with the aim of improving model performance. These runs were explored to determine if OSA residual patterns for fit to age composition data in runs 29-29C could be resolved. Runs 29F-29F5 implemented a logistic normal age composition model that estimates an

additional weighting parameter for each index and treated zero observations as missing. Run 29H was an exploratory run that implemented a dirichlet-multinomial age composition model to directly estimate effective sample size. These runs were fit to split Albatross and Bigelow indices (i.e. Bigelow years not calibrated to Albatross units) with aggregate index data in biomass units as in runs 29A-29C.

Run 29F only converged when a wider range of selectivities were fixed at 1 compared to run 29B, and included age 11+ for the Albatross fall index. F estimates tended to be lower than estimates from run 27 and other split runs (29-29C), and had a slightly different trajectory although most major peaks/valleys were generally captured. R estimates tracked the estimates from other runs fairly well and tended to be on the higher side of the range. SSB estimates were similar or slightly higher prior to 2000 but fell between the estimates for run 29 (on the higher end) and run 27 (on the lower end). CVs around estimates of F were more variable over time and were higher for estimates of R. OSA residuals for the fleet and aggregate spring indices were similarly or slightly more normally distributed and fall indices were slightly less normally distributed than in run 29F compared to 29B. Age composition OSA residuals were more normally distributed for the fleet and all indices, although their magnitude was a bit larger. Some years had residual patterns where residuals for all or most ages were consistently positively or negatively biased.

Runs 29F1-29F5 varied starting selectivity estimates and run 29F2 explored the inclusion of a selectivity random effect to try to improve the estimation of age 11+ selectivity for the Albatross fall index. Runs 29F1-29F5 had larger AIC values than run 29F so they were not considered an improvement on this prior run based on this statistic alone. However, runs 29F2, 29F4, and 29F5 had other model improvements that qualified them for further consideration.

Run 29F1 used the selectivity estimates from run 29B as starting estimates for the run, and initially tried to estimate selectivity at older ages freely in a preliminary run. However, the final version of this run still required that age 11+ selectivity for the Albatross fall index was fixed at 1 in order to converge. This run had minor differences in selectivity estimates for both spring and the Bigelow fall index but had little effect on estimates for the Albatross fall index.

Run 29F2 used the selectivity estimates from run 29B as starting estimates for the run and implemented a selectivity random effect for the Albatross fall index. When this initial run failed to converge, the Albatross fall age 11+ selectivity was fixed at the starting estimate (0.5783887) and the revised run converged. Although the AIC value was slightly larger than in run 29F, the OSA residuals for fit to the aggregate fleet, and both

spring and fall Albatross indices were more normally distributed, indicating that this model more appropriately fits to this data. This run was one of three candidate runs (see Section 4.12 for full details).

Run 29F3 used the selectivity estimates from run 29B and fixed age 11+ selectivity for the Albatross fall index at this starting estimate (0.5783887). Estimates of selectivity-at-age followed similar patterns to run 29F but were generally smaller in magnitude for ages that were not fully selected in run 29F-3. Other diagnostics were similar to those in run 29F.

Run 29F4 reverted to using 0.5 as starting estimates for selectivity-at-age and used an initial run to identify a single age for each index to fix at full selectivity. This resulted in a final run that fixed Albatross spring age 6, Bigelow spring age 5, Albatross fall age 4, and Bigelow fall age 3 selectivity at 1. This run successfully converged while freely estimating the selectivity for Albatross fall age 11+, but this estimate was near 1 and had wide confidence bounds that spanned from 0 to 1. Although the AIC value was larger than for run 29F (and thus less preferred based on this metric), the free estimation of selectivity for the Albatross fall plus group was preferable as it more appropriately represents the uncertainty in this estimate than in runs that fixed this parameter. Uncertainty in selectivity estimates for Albatross spring age 5, Bigelow spring age 4, and Bigelow fall age 4 were also highlighted in this run with confidence bounds around estimates spanning from near 0 to near 1. This run was one of three candidate runs (see Section 4.12 for full details).

Run 29F5 was specified as in run 29F4 but explored two approaches to freely estimate the Albatross fall age 11+ selectivity at value farther from 1 (full selectivity). A preliminary run freely estimated Albatross fall selectivity-at-age but forced the estimate for age 10 and 11+ to match, resulting in an estimate near full selectivity (0.9781860). Because this did not lower the estimate of selectivity for Albatross fall age 11+, the full run instead implemented an AR1 random effect on age for this index (age-varying rather than time-varying). This change resulted in slightly lower selectivity estimates than in run 29F4 for the Albatross fall index except for age 4 which was fixed at full selectivity. The OSA residuals for fit to the aggregate fleet and index data were similarly or slightly more normally distributed than in run 29F4, and the Albatross fall residuals in particular were more normally distributed. Mohn's rho and AIC values were larger than those for run 29F4. This run was one of three candidate runs (see Section 4.12 for full details).

A single run (29H) explored a dirichlet-multinomial age composition likelihood model. This run had difficulty calculating the OSA residuals so they were not used to compare with other models. The AIC value was not comparable to other runs due to the change

in likelihood structure, but Mohn's rho values were comparable, with SSB and Fbar rho values smaller than in runs 29F, 29F2 and 29F4, and R rho value larger than in these three runs. Mohn's rho values for run 29 were always smaller than those for run 29F5.

4.11 Environmental Covariate Runs

Work conducted under ToR 1 (see section 1) identified the following potential drivers of recruitment (R) and survey catchability (q) and were thus explored in WHAM runs that linked environmental covariates to these stock dynamics: sea surface temperature anomalies (SST anomaly), bottom temperature anomalies (BT anomaly) and the North Atlantic (NAO), and Atlantic Multidecadal Oscillations (AMO). A preliminary run of each environmentally-linked model was conducted to fit to the environmental covariate data without an effect on stock dynamics specified to allow comparisons via AIC to models with the covariate effect specified (runs 39, 41, 42, 43, 44, 46, 47, 48, 49, and 50). OSA residuals for fit to the environmental covariates were underdispersed for two of these runs when no effect was specified (42 fit to NAO and 48 fit to AMO) and estimated environmental covariates had much narrower distributions than the observed data (Figure 4.6). These results indicate a modeling error that may also impact AIC calculations and thus make this model inappropriate for comparison with other models. Although runs 42 and 48 were most strongly impacted by the modeling error, its existence lowers our confidence that AIC values are accurate for other runs that fit to environmental covariate data without specifying a covariate effect. Within the timeframe of this research track assessment we were unable to resolve the underlying modeling issue and could not make confident comparisons between runs that did and did not specify an environment covariate effect so our advice is to exclude environmental covariates from candidate models at this time. However, we outline some general conclusions drawn from the runs that did specify an environmental covariate effect that may warrant further consideration in future analyses.

Both random walk and AR1 environmental processes were explored for covariates that affected either recruitment or catchability. In most cases there were few or no differences between models based on what process was implemented, but the choice of environmental process did impact the sign and magnitude of bottom temperature effects on catchability (runs 41A and 47A). Catchability generally had a positive relationship with bottom temperature (except for the Bigelow spring index when an AR1 process was assumed). This conclusion aligns with the results of preliminary analyses from ToR1 (see section 1) which found decreasing catchability as plaice moved into deeper, colder water and highlights these runs as key models to revisit in future analyses. Changes in the magnitude or sign of the environmental effect could indicate a misspecified environmental process in one of these models, but could also be attributed to a misspecified effect on one or more of the indices. We recommend exploring the

consequences of specifying an environmental effect on catchability for a subset of the available indices in future analyses to evaluate the later source of misspecification. Furthermore where runs do not have variable performance based on the environmental process model, there was a working group recommendation to implement an AR1 process because the variance of projections asymptotes for this process rather than going to infinity as for random walk processes. This is not expected to influence model results but has consequences for model projections.

In contrast to the models with environmental effects on catchability, runs that affected recruitment generally had a negative relationship (e.g. run 46 estimated larger recruitments as sea surface temperature anomalies became more negative). This result was in conflict with the preliminary analyses from ToR1 (see section 1) that suggested increasing recruitment as sea surface temperature increased. This conflict further justified the selection of candidate models without environmental effects on recruitment in this research track, but this conclusion should be reassessed in future analyses (after addressing the underlying modeling errors) since the exploratory analyses from ToR1 strongly suggested an environmental relationship with recruitment.

To ensure that future model explorations with environmental covariates are comparable via AIC, we also recommend including all available covariates in each model but only specify links for those that are being actively evaluated (i.e. similar to the current approach of fitting to the covariates without specifying an effect, but fit to multiple covariate data sets at once).

4.12 Candidate Models

Three models emerged as candidates (29F2, 29F4, and 29F5) and are compared in greater detail here. All three candidate models fit to aggregate (kg/tow) and age composition (abundance) data for four indices and a single fleet. NEFSC bottom trawl survey data was split into separate Albatross and Bigelow indices for both the spring and fall, and Bigelow data was uncalibrated (i.e. no calibration to Albatross units was used) from 2009-2019. Natural mortality was fixed at 0.3 following a revision from 0.2 in the prior VAST assessment and a constant maturity-at-age schedule was implemented in all three runs. No stock-recruit relationship was estimated and recruitment was instead assumed to be random about an estimated mean. Random effects were implemented for all numbers-at-age, to allow variable survival for each age class and year. Improved fit to catch data and a lower AIC score were associated with run 16 which implemented two selectivity blocks to account for differences between historic and contemporary fishing patterns, but this approach did not capture shorter-term

effects of regulatory changes on selectivity. Candidate models built on this improvement by implementing a selectivity random effect for the fleet, allowing for time-varying selectivity. All candidate runs assumed logistic selectivity for the fleet and estimated age-specific selectivity for the four indices. A logistic-normal age composition model was implemented for each of the candidate runs.

4.12.1 Selectivity

All three candidate runs have a similar model configuration but sought to address uncertainty in age 11+ selectivity for the Albatross fall index in different ways, and consequently had minor differences in selectivity estimates (Figure 4.7). Run 29F2 implemented a selectivity random effect for only the Albatross fall index and fixed age 11+ selectivity at the estimated value from run 29B (0.5783887). Selectivity was fixed at 1 for ages 5 and 6 in the Albatross spring index, age 5 in the Bigelow spring index, age 4 in the Albatross fall index, and ages 3 and 4 in the Bigelow fall index, with starting estimates set to estimated values from run 29B except for Albatross spring age 4 and Bigelow fall age 5 selectivies which had starting estimates set to 0.5.

Run 29F4 only fixed a single age at full selectivity (age 6 in Albatross spring, age 5 in Bigelow spring, age 4 in Albatross fall, and age 3 in Bigelow fall), and set index selectivity starting estimates for all other ages to 0.5. No selectivity random effects were included for the indices. Selectivity for Albatross fall age 11+ was freely estimated in this run but had an estimate near 1 with large confidence bounds (Figure 4.7).

Run 29F5 was specified identically to run 29F4 but modeled age-specific selectivity for the Albatross IV fall index as an ar1 random effect (not time varying, just age varying), aside from the age fixed at full selection. Treating age 11+ as a random effect in this way resulted in a lower estimate than in run 29F4 (Figure 4.7).

4.12.2 Convergence

All three candidate runs met first and second order convergence criteria.

4.12.3 AIC

AIC scores for run 29F2 and 29F4 were within +/-2 of each other so these runs should be considered equivalent, but of the two 29F4 had the slightly lower score. Other diagnostics should inform the selection between the two. AIC for run 29F5 was not comparable to the other candidate runs due to a difference in likelihood structure so

other diagnostics should be used to select between this model and the other candidates.

4.12.4 Mohn's rho

None of the candidate runs had strong retrospective patterns, and consequently there were only minor differences in Mohn's rho values between these runs (Figure 4.8). Run 29F5 had the largest Mohn's rho values out of the three runs, and run 29F4 had the smallest values for R and SSB while run 29F2 had the smallest Mohn's rho value for Fbar.

4.12.5 One-step-ahead residuals

The distribution of OSA residuals varied slightly between candidate runs for fit to both aggregate and age composition data, but were generally normally distributed (Figure 4.9). Runs 29F4 had a slightly different distribution of OSA residuals for fit to the Albatross fall index than run 29F2 and 29F5 which respectively fixed or estimated age 11+ selectivity at a lower value.

4.12.6 MASE

Mean absolute scaled error (MASE) scores are a measure of model and index prediction skill. On average the Bigelow fall index was more accurately predicted than the spring index, but confidence bounds around the fall MASE scores were larger (Figure 4.10).

MASE scores for all three candidate runs decreased as the prediction horizon increased from 1-3 years (typical management horizons) and started to increase for Bigelow fall horizons greater than 4 years (Figure 4.11). Generally we would expect prediction skill to decrease as prediction horizon increases, but it is possible to see the opposite trend for models that appropriately describe the data or when there is a trend in the index itself. In such cases longer horizons provide more data for use in predictions and are thus improved (smaller MASE scores) over shorter horizons that use less data for predictions.

On average runs 29F4 and 29F5 had slightly lower MASE scores than run 29F2 over a typical management horizon (1-3 years), indicating slightly improved predictive skill for these two candidates (Figure 4.12). However, the differences are sufficiently small that this statistic alone does not strongly point to the selection of one candidate over another.

4.12.7 Self-tests

One hundred self-tests were conducted for each of the candidate models by first using the candidate model to generate one hundred aggregate catch and index data, catch and index age composition data, and observation error data sets. The candidate models were then re-fit to each of the data sets generated in the first step. Convergence rate and relative errors for SSB, F, recruitment, and catch were used to compare self-test performance between candidates.

Run 29F2 had the lowest convergence rate (66.7%) of the three candidates, with runs 29F4 and 29F5 always converging. Median relative errors for SSB and recruitment were greater than 1, indicating the tendency to overestimate these values (Figures 4.13, 4.14) . In particular, there was more variability in the SSB relative error for runs 29F4 and 29F5 towards the end of the time series than was observed in run 29F2. The relative error for recruitment was larger in magnitude and had larger interannual differences than other relative errors. Runs 29F4 and 29F5 had median relative recruitment errors closer to 1 than run 29F2, but run 29F4 had a wider spread than run 29F5.

Catch relative errors were very slightly larger than 1 so there is a minor tendency to overestimate catch but the scale of this overestimation was much smaller than for SSB or recruitment (Figure 4.15). Of the four relative error metrics, catch relative error had the smallest magnitude, very low interannual variation, and had median values closest to 1. Relative errors in the last 10-15 years of the time series had very little differences between the middle 50% and 80% of catch relative errors across simulations and were very close to 1.

Relative error for fishing mortality was the only metric with median relative errors smaller than 1, indicating the tendency to underestimate F in self-tests (Figure 4.16). This pattern was particularly apparent over the last 10 years of the time series for all three candidate models, where median values were mostly less than 1.

4.12.8 Model estimates

All three candidate models estimated similar trends in R, R and SSB and had similar trends in CVs around these estimates but the scale varied slightly between models (Figure 4.17). Runs 29F4 and 29F5 had very similar estimates and trends in CVs, but had generally higher F estimates, slightly higher R estimates, and slightly lower SSB estimates compared to run 29F2. This corresponded to similar CVs around F estimates in all runs, but run 29F2 had slightly higher CVs around R estimates early in the time

series and consistently higher CVs around SSB estimates compared to runs 29F4 and 29F5.

4.15 References

Akaike, H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control, 19(6):716-723.

Carvalho F, Winker H, Courtney D, Kapur M, Kell L, Cardinale M, Schirripa M, Kitakado T, Yemane D, Piner KR, Maunder MN. 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240:105959.

Fisch N, Camp E, Shertzer K, Ahrens R. 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fish. Res. 243:106069.

Kell LT, 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 J. Mar. Sci. 78(6):2244-2255.

Legault CM, Restrepo VR. 1999. A flexible forward age-structured assessment program. ICCAT Collect. Vol. Sci. Pap. 49(2):246–253. SCRS/98/058. [Available from website.]

Mohn R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56:473–488.

NEFSC (Northeast Fisheries Science Center). 2008. Assessment of 19 Northeast Groundfish Stocks through 2007: Report of the 3rd Groundfish Assessment Review Meeting (GARM III), Northeast Fisheries Science Center, Woods Hole, Massachusetts, August 4-8, 2008. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 08-15; 884 p + xvii. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026.

NEFSC (Northeast Fisheries Science Center). 2012. Assessment or Data Updates of 13 Northeast Groundfish Stocks through 2010. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 12-06; 789 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/

NEFSC (Northeast Fisheries Science Center). 2015. Operational Assessment of 20 Northeast Groundfish Stocks, Updated Through 2014. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 15-24; 251 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/publications/

NEFSC (Northeast Fisheries Science Center). 2017. Operational Assessment of 19 Northeast Groundfish Stocks, Updated Through 2016. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 17-17; 259 p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/publications/

NEFSC (Northeast Fisheries Science Center). 2019. Stock Assessment Update of 14 Northeast Groundfish Stocks Through 2018. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 19-XXXX; ??p. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/

Stock BC, Miller TJ. 2021. The Woods Hole Assessment Model (WHAM): A general state-space assessment framework that incorporates time-and age-varying processes via random effects and links to environmental covariates. Fish. Res. 240:105967.

Thygesen UH, Albertsen CM, Berg CW, Kristensen K, Nielsen A. 2017. Validation of ecological state space models using the Laplace approximation. Environ. Ecol. Stat. 24(2):317-339.

4.16 Tables and Figures

Table 4.1: Description, AIC, and Mohn's rho values for all runs that met convergence criteria, highlighting models that are comparable via AIC in similar colors.

Description	Run	AIC	Rho_R	Rho_SSB	Rho_Fbar
Run fit to revised data (ages 1-11+), updated maturity schedule, and updated natural mortality assumptions (0.3), selectivity as specified for input data file.	9	5925.4	-0.5181	0.0161	-0.0343

Run fit to revised data (ages 1-11+) and updated maturity schedule but revert to the natural mortality assumptions used in the VPA (0.2), selectivity as specified for input data file.	10	5606.6	-0.491	0.2305	-0.2373
Run fit to revised data (ages 1-11+) but revert to the maturity and natural mortality assumptions used in the VPA (0.2), selectivity as specified for input data file.	11	5606.6	-0.491	0.2334	-0.2373
Logistic selectivity for the fleet and age-specific selectivity for spring and fall NEFSC indices.	13	4259.9	0.3916	0.0448	-0.0467
Explore logistic selectivity for the NEFSC fall index.	14	4276.1	0.4082	0.0796	-0.0851
Explore two selectivity blocks (1980-1999 and 2000-2019)	16	4253.5	0.3891	0.0529	-0.0452
Two selectivity blocks (1980-1999 and 2000-2019), with recruitment random effect specified.	17	4351.2	0.4974	0.054	-0.0485
Two selectivity blocks (1980-1999 and 2000-2019), logistic-normal age composition likelihood, NEFSC selectivity switched from age-specific (as in run 17) to logistic.	19	-5072.7	0.265	0.0457	-0.038
Two selectivity blocks (1980-1999 and 2000-2019) with effective sample size changed to match that in ASAP run 12	22	3026.9	0.4008	0.0099	-0.0032
Run with iid selectivity random effects for the fleet and all indices.	23	4168.0	0.3391	0.0488	-0.0450
Run with ar1_y selectivity random effects for the fleet and all indices.	24	4114.6	0.7195	0.3203	-0.2813
Run with only recruitment random effects and iid selectivity random effects for the fleet and all indices.	25	4259.4	0.5419	0.0606	-0.0539

Run with only recruitment random effects and ar1_y process.	26	4257.3	0.5852	0.0587	-0.0514
Full state space model fit to combined NEFSC indices, Albatross and abundance units	27	4210.1	0.3765	0.0361	-0.0331
Combined, Albatross & biomass units	27A	4221.0	0.2475	-0.0379	0.0890
Run with only recruitment random effects, fit to extended catch time series that extends to 1960 and several additional years of aggregate survey data (without age composition data).	28	4393.1	0.3780	0.0555	-0.0409
Split Albatross/Bigelow time series with multinomial age comp, Albatross units, abundance units & selectivity random effects for Albatross spring index	29	4183.3	-0.0221	0.0071	-0.0143
Split, Bigelow & biomass units	29A	4209.7	0.0223	0.0037	0.0061
Same as run 29B but ESS changed to reflect ASAP run 51a expectations	29B-1	4129.1	0.0104	0.0029	0.0075
Split Albatross/Bigelow time series with multinomial age comp, Bigelow units, biomass units & no selectivity random effects	29B	4208.0	0.0212	0.0040	0.0059
Split, Albatross & biomass units	29C	4217.6	0.0274	0.0033	0.0064
Split Albatross/Bigelow time series with logistic normal age comp, Bigelow units & biomass units	29F	-5493.5	-0.0778	-0.0243	0.0274
Rerun 29F but use selectivity estimates from run 29B as starting estimates, still need to fix Albatross fall age 11+ selectivity at 1	29F-1	-5488.1	-0.0753	-0.0275	0.0353
Rerun 29F but use selectivity estimates from run 29B as starting estimates , fix Albatross fall age	29F-2	-5480.6	-0.0825	-0.0311	0.0319

11+ selectivity at 29B estimate & apply selectivity random effect to Albatross fall index					
Rerun 29F but use selectivity estimates from run 29B as starting estimates , fix Albatross fall age 11+ selectivity at 29B estimate	29F-3	-5469.3	-0.0842	-0.0311	0.0323
Rerun 29F but fix only 1 age at full selectivity for each index based on initial run that used 0.5 as starting estimate	29F-4	-5482.3	-0.0724	-0.0278	0.0370
Implementation like run 29F-4, but also implement an ar1 random effect on age for the Albatross fall index.	29F-5	-5453.0	-0.1378	-0.0540	0.0607
Run similar to 29F with Dirichlet multinomial age composition model	29H	8435.2	-0.1143	0.0159	-0.0017
Run with random effects for all numbers-at-age fit to NEFSC and MADMF spring and fall indices with selectivity random effect only implemented for the fleet.	30	6847.0	-0.0363	0.0419	-0.0590
Run with random effects for all numbers-at-age fit to NEFSC and MENH spring and fall indices.	31	5447.3	0.0710	0.0315	-0.0545
Run with only recruitment random effects fit to NEFSC spring and fall indices with selectivity random effects and an LPUE index with selectivity that mirrors the fleet.	32	4289.0	0.5709	0.0576	-0.0491
Run fit to NEFSC spring and fall indices with selectivity random effects, and an LPUE index with selectivity mirroring the fleet and with a catchability random effect applied.	32A	4286.1	NA	NA	NA
Run with random effects for all numbers-at-age fit to NEFSC,	33	8252.2	-0.0557	0.0507	-0.0704

MADMF and MENH spring and fall indices with a selectivity random effect only applied to the fleet.					
Run with only recruitment random effects fit to NEFSC, MENH, and LPUE indices with selectivity random effects on the fleet and all but the LPUE index, and LPUE selectivity mirrored to the fleet.	34A	5644.7	0.1513	0.0228	-0.0414
Run with only recruitment random effects fit to NEFSC, MADMF, and LPUE indices with selectivity random effects on only the fleet and LPUE selectivity mirrored to the fleet.	35A	7253.7	-0.0268	0.0004	-0.0277
Full state-space model fit to updated VAST run & logistic selectivity for the fleet and both spring and fall indices	37E	5730.4	-0.0133	0.0875	0.0486
Run 29B but fit to SST anomaly with no effect specified	*39	NA	0.0212	0.0040	0.0059
Same as run 39 but has SST anomaly effect on R with a 1 year lag	39A	4270.5	-0.0820	0.0185	-0.0177
Run 29B but fit to BT anomaly with no effect specified	*41	NA	0.0212	0.0040	0.0059
Same as run 41 but has BT anomaly effect on catchability for all indices	41A	4249.6	-0.0807	-0.0435	0.0414
Run 29B but fit to NAO data with no effect specified	*42	NA	0.0209	0.0035	0.0046
Same as run 42 but has NAO effect on R with a 1 year lag	42A	4265.8	0.0371	0.0118	-0.0007
Same as run 42 but has NAO effect on R with a 2 year lag	***42B	4265.9	0.0131	0.0033	0.0080

Run 29B but fit to AMO data with no effect specified	*43	NA	0.0209	0.0035	0.0046
Same as run 43 but has AMO effect on R with a 1 year lag	43A	4178.1	0.0644	0.0238	-0.0145
Run 29B but fit to BT anomaly data with no effect specified	*44	NA	0.0212	0.0040	0.0059
Same as run 44 but has BT anomaly effect on R with a 1 year lag	44A	4265.0	0.0082	0.0356	-0.0279
Identical to run 39 with fit to SST anomaly data without an effect specified, but implement an ar1 process	**46	NA	0.0212	0.0040	0.0059
Identical to run 39A with SST anomaly effect on R, but implement an ar1 process	46A	4264.9	-0.0697	0.0222	-0.0205
Identical to run 41 with fit to BT anomaly without an effect specified, but implement an ar1 process	**47	NA	0.0212	0.0040	0.0059
Identical to run 41A with BT anomaly effect on q but implement an ar1 process	47A	4257.8	-0.0028	-0.0325	0.0419
Identical to run 42 fit to NAO data with no effect specified, but implement an ar1 process	**48	NA	0.0209	0.0035	0.0046
Identical to run 42A with NAO effect on R with 1 year lag, but implement an ar1 process	48A	4257.2	0.0348	0.0072	0.0028
Identical to run 42A with NAO effect on R with 2 year lag, but implement an ar1 process	***48B	4257.8	-0.0174	-0.0116	0.0198
Identical to run 43 fit to AMO data with no effect specified, but implement an ar1 process	**49	NA	0.0209	0.0035	0.0046

Identical to run 43A with AMO effect on R, but implement an ar1 process	49A	4174.2	0.0617	0.0226	-0.0134
Identical to run 44 fit to BT anomaly with no effect specified, but implement an ar1 process	**50	NA	0.0212	0.0040	0.0059
Identical to run 44A with BT anomaly effect on R, but implement an ar1 process	50A	4265.9	0.0085	0.0357	-0.0280

* Identical model specification to run 29B but fit to an environmental covariate with a random walk process so theoretically comparable via AIC to runs that link these covariates to recruitment or catchability, but AIC not reported here due to modeling error.

** Identical model specification to run 29B, but fit to an environmental covariate with an ar1 process so theoretically comparable via AIC to runs that link these covariates to recruitment or catchability, but AIC not reported here due to modeling error.

*** Run 42B and 48B fit to one year less environmental data than runs 42 and 48 respectively due to a 2 year lagged effect, so they are not technically comparable via AIC, but the difference in data fit in each is sufficiently small that general comparisons were made here.

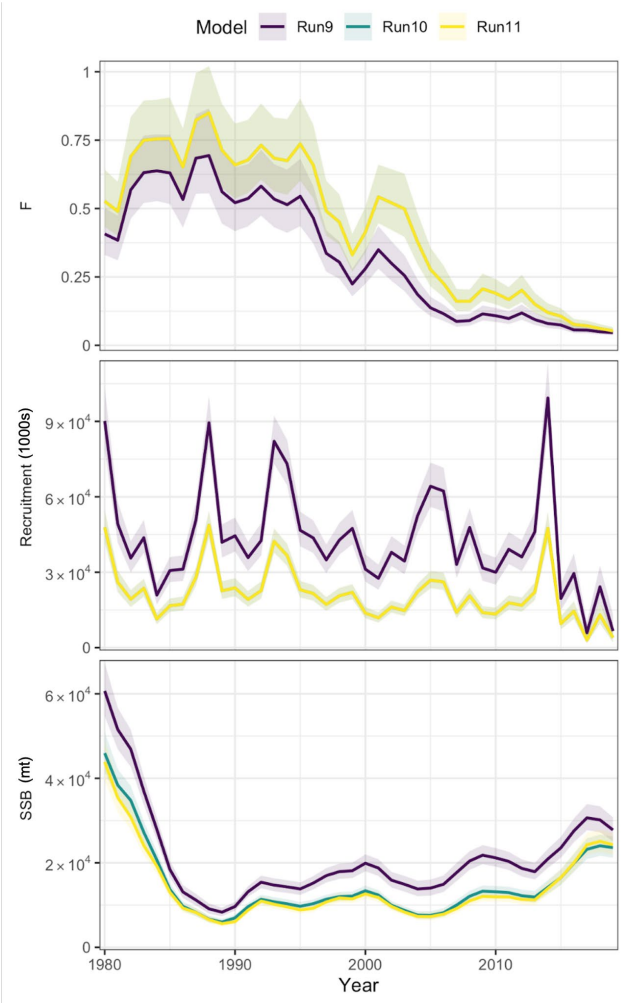

Figure 4.1: Model estimates of spawning stock biomass (SSB), fully-selected fishing mortality (F), and recruitment for runs where all data, natural mortality (M), and maturity assumptions were updated (run 9, purple line), only data and maturity were updated (M remained at 0.2; run 10, green line), and a run where only the data were updated but M and maturity assumptions remained identical to those in the VPA (run 11, yellow line).

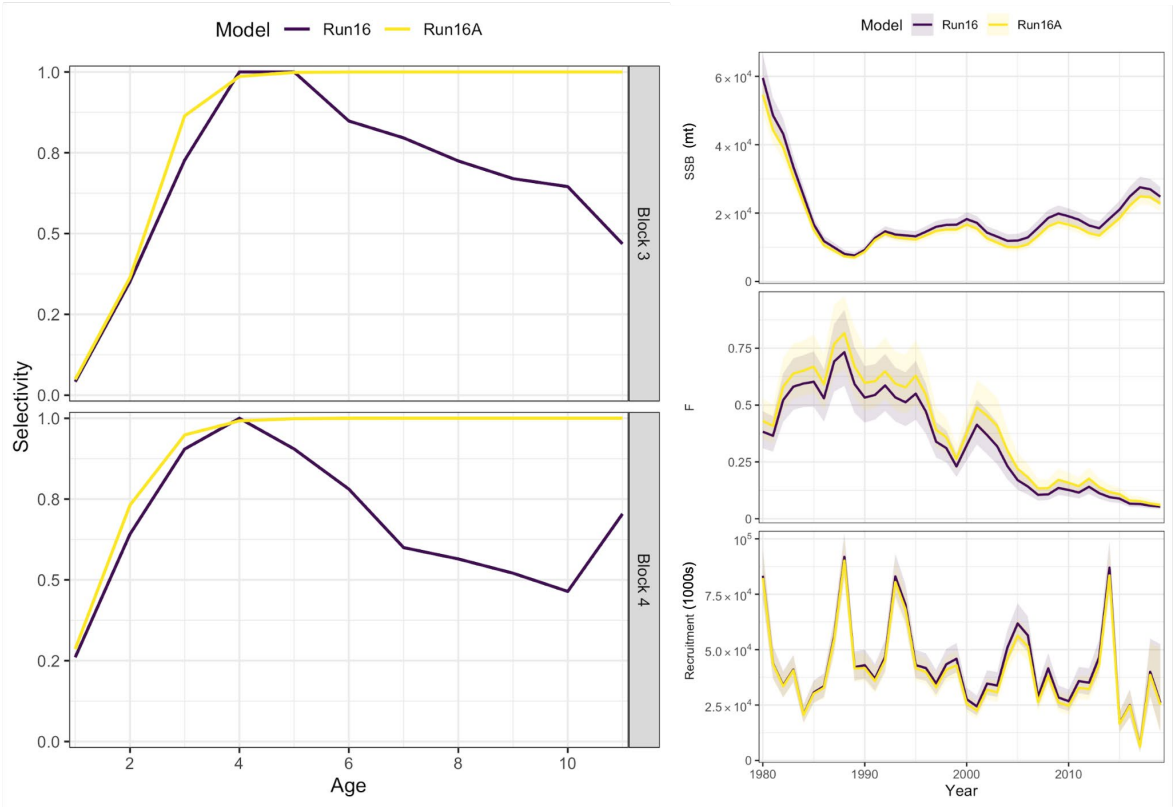

Figure 4.2: Selectivity (left) for NEFSC spring (block 3) and fall (block 4) indices and model predictions of SSB, F, and recruitment (right). Run 16 assumed age-specific selectivity for both spring and fall NEFSC indices while run 16A assumed logistic selectivity for these indices.

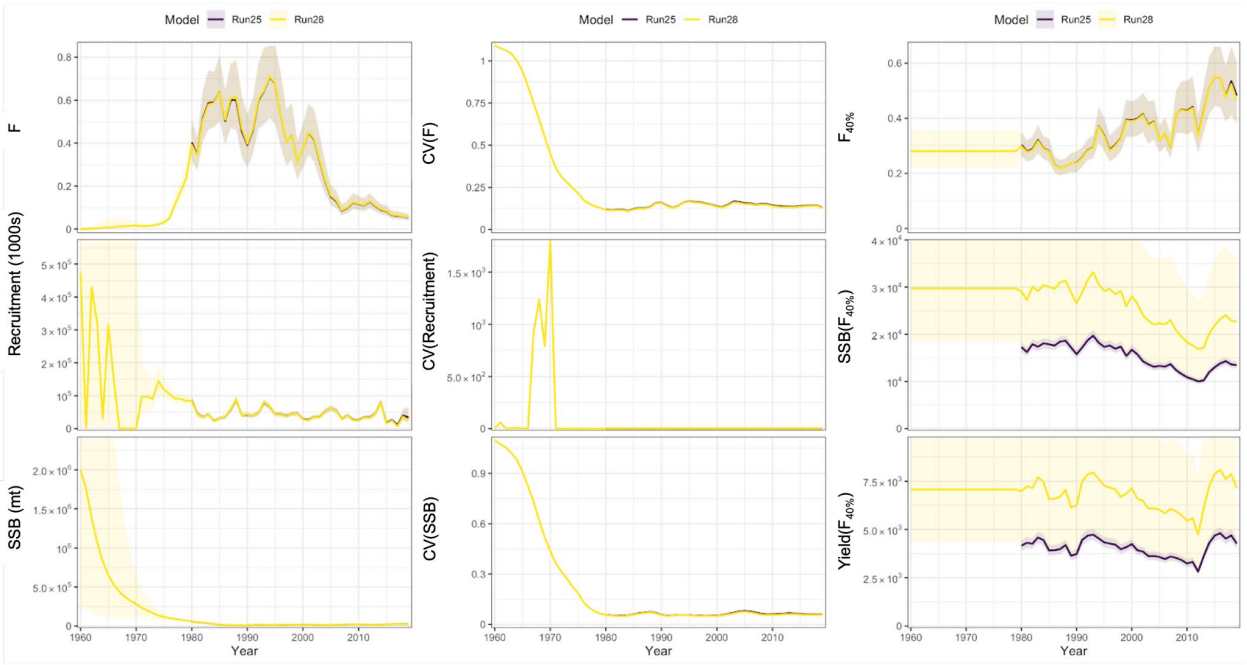

Figure 4.3: Model estimates of spawning stock biomass (SSB), fishing mortality (F) and recruitment (left), corresponding CVs around these estimates (center) and F40%, SSB at F40%, and Yield at F40% reference points (right). Both run 25 and 28 implemented recruitment and selectivity random effects, with run 28 fit to catch data beginning in 1960 rather than in 1980 as in run 25.

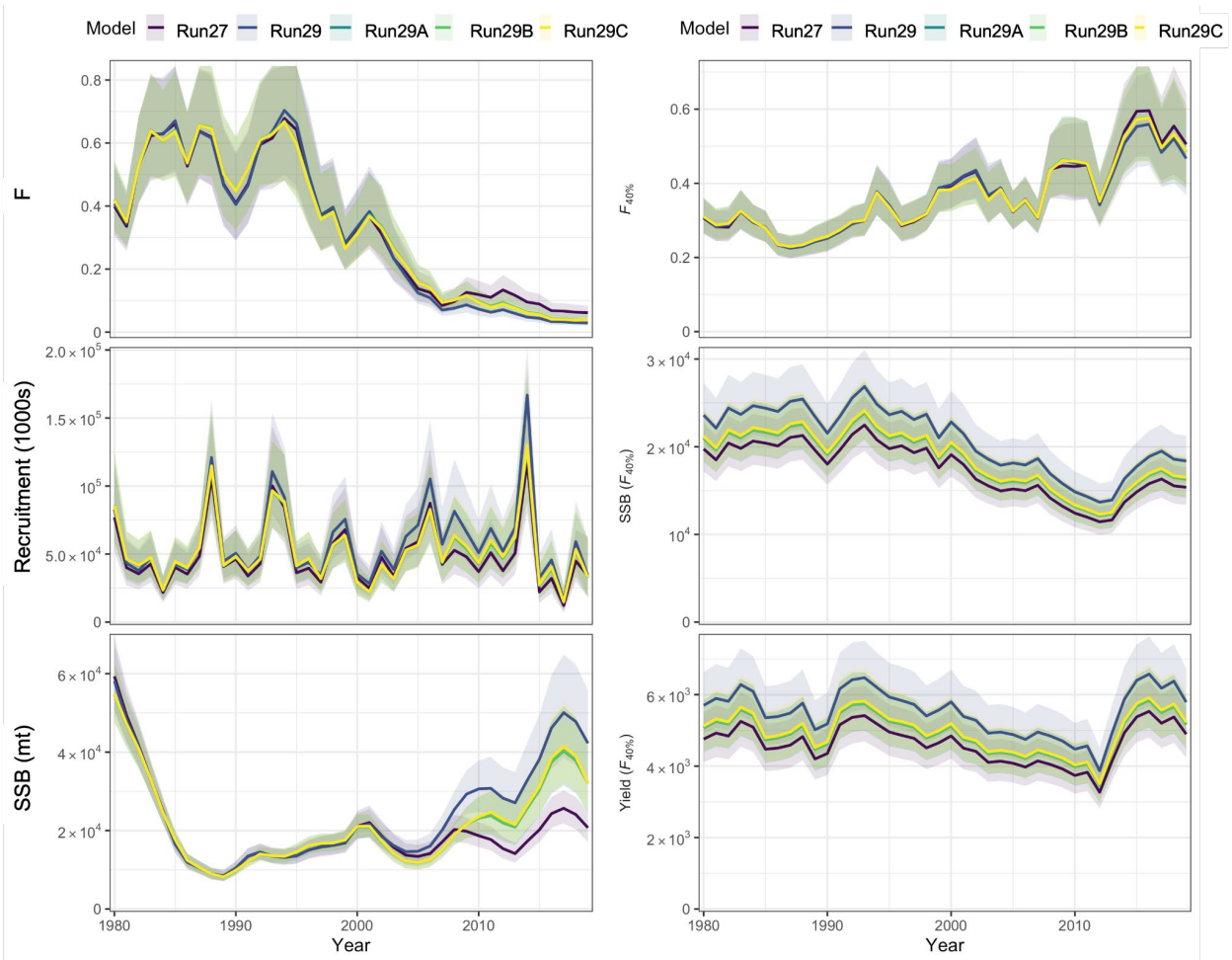

Figure 4.4: Model estimates of spawning stock biomass (SSB), fishing mortality (F) and recruitment (left) and reference points F40%, SSB at F40%, and Yield at F40% (right). Run 27 implemented a full state-space model with iid selectivity random effects fit to the full NEFSC spring and fall indices (1980-2019), while runs 29-29C were fit to split Albatross (1980-2008) and Bigelow (2009-2019) spring and fall indices. Run 29 fit to age composition in numbers-at-age and Albatross survey units, runs 29A-B fit to biomass age composition and Bigelow survey units from 2009-2019 but 29B excluded index selectivity random effects, and run 29C fit to biomass age composition but used Albatross survey units from 1980-2008 and 2009-2019.

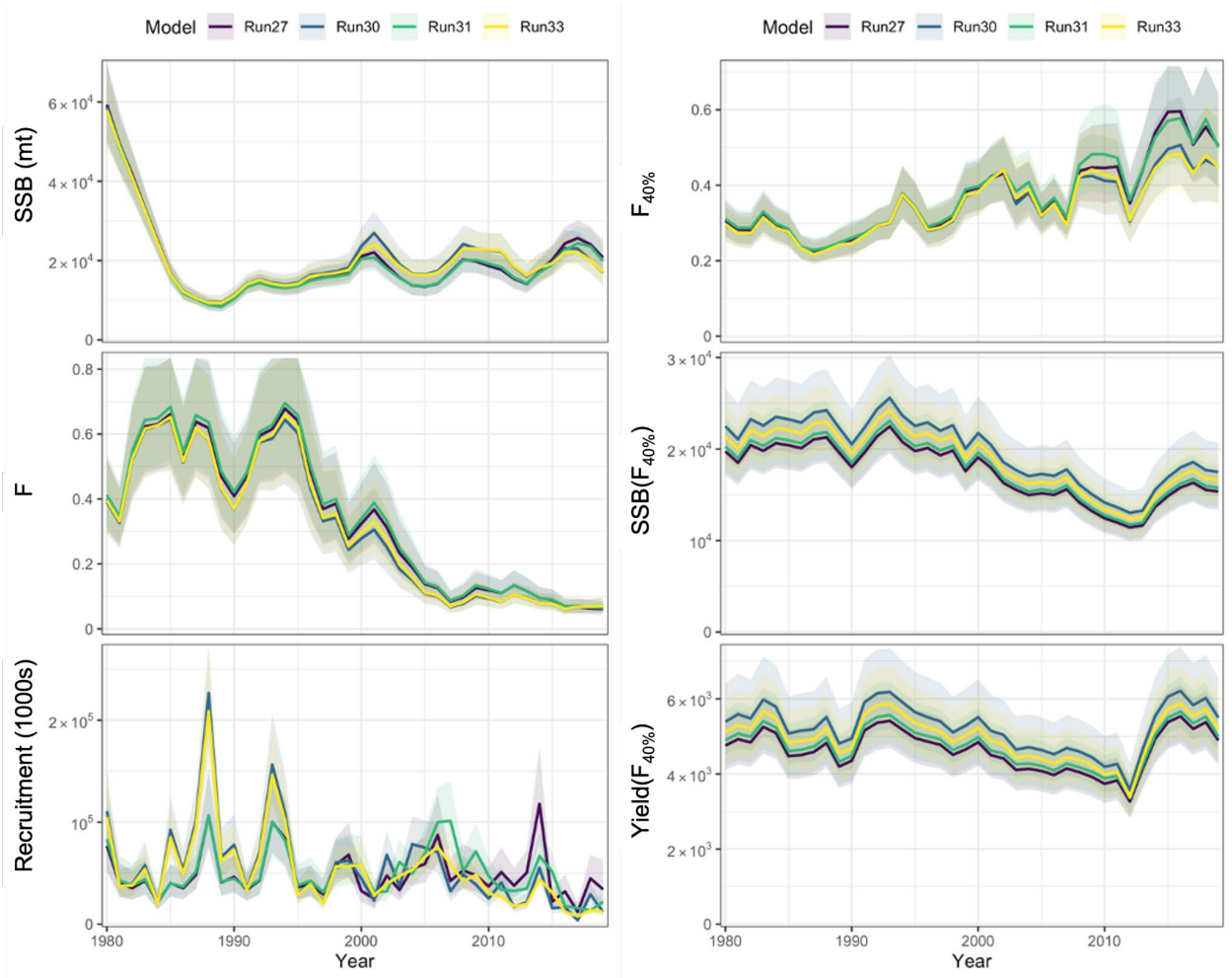

Figure 4.5: Model estimates of spawning stock biomass (SSB), fishing mortality (F) and recruitment (left) and reference points F40%, SSB at F40%, and Yield at F40% (right). Run 27 implemented a full state-space model with iid selectivity random effects fit to the full NEFSC spring and fall indices, run 30 fit a full state-space model to NEFSC and MADMF indices without selectivity random effects, run 31 fit a full state-space model to NEFSC and ME-NH indices with iid selectivity random effects, and run 33 fit a full state-space model to NEFSC, MADMF, and ME-NH indices without selectivity random effects.

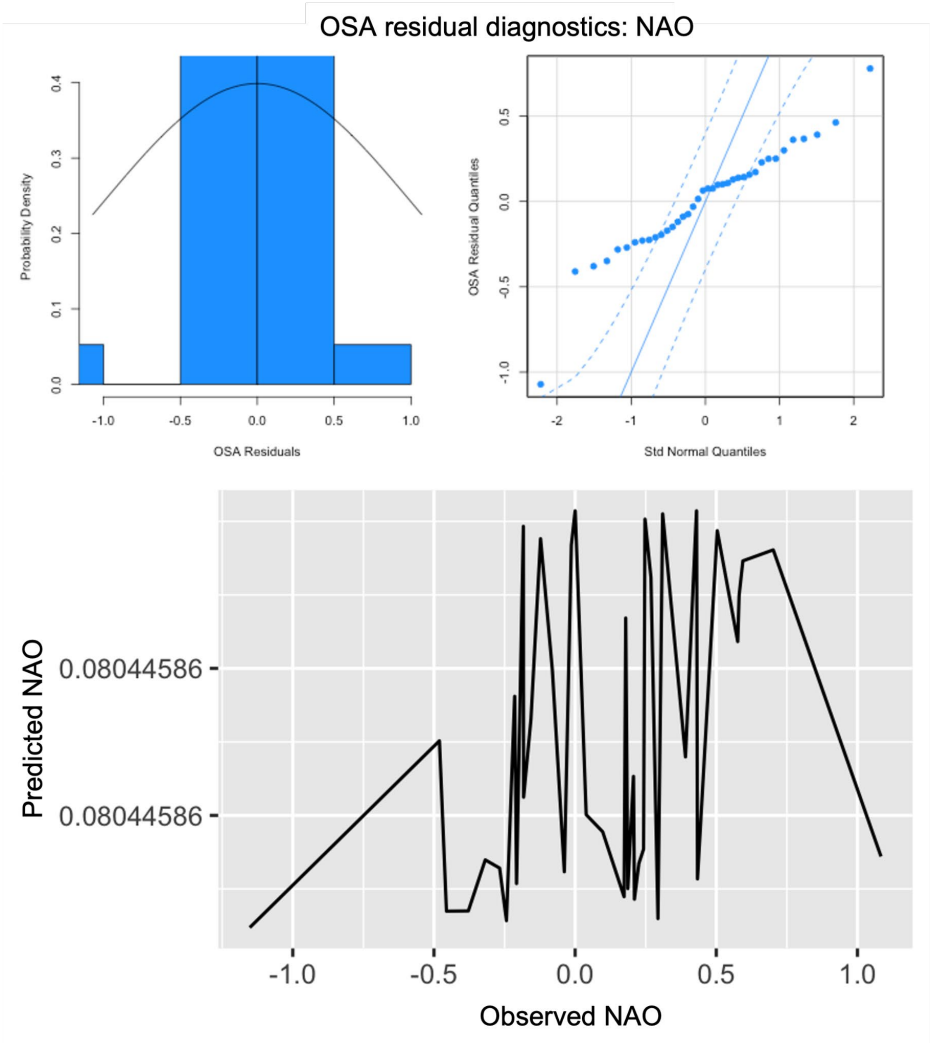

Figure 4.6: Example of overdispersed one-step ahead (OSA) residuals for fit to a North Atlantic Oscillation (NAO) covariate (top) and comparison of model predicted values for this covariate vs. observed values for this covariate (bottom). Both results indicate a modeling error in WHAM when an model is fit to an environmental covariate without an effect on stock dynamics specified (i.e. to establish a base model that is comparable via AIC to runs with an effect specified) that was not resolved in this research track due to time constraints.

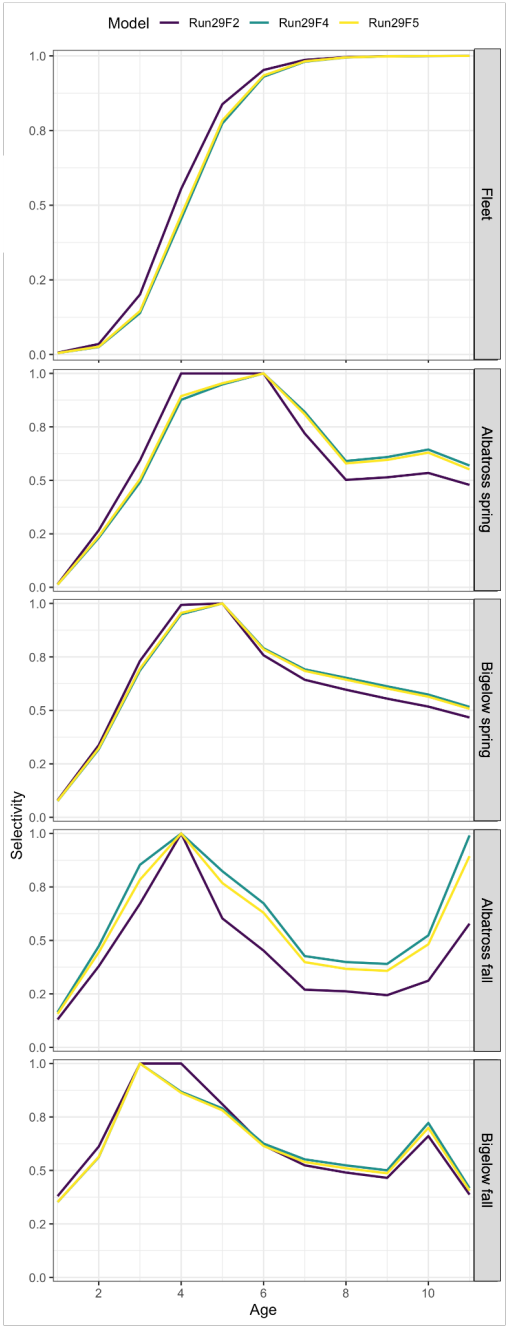

Figure 4.7: Selectivity estimates for candidate runs 29F2 (purple), 29F4 (green) and 29F5 (yellow) for the fleet (logistic selectivity) and four indices (selectivity-at-age).

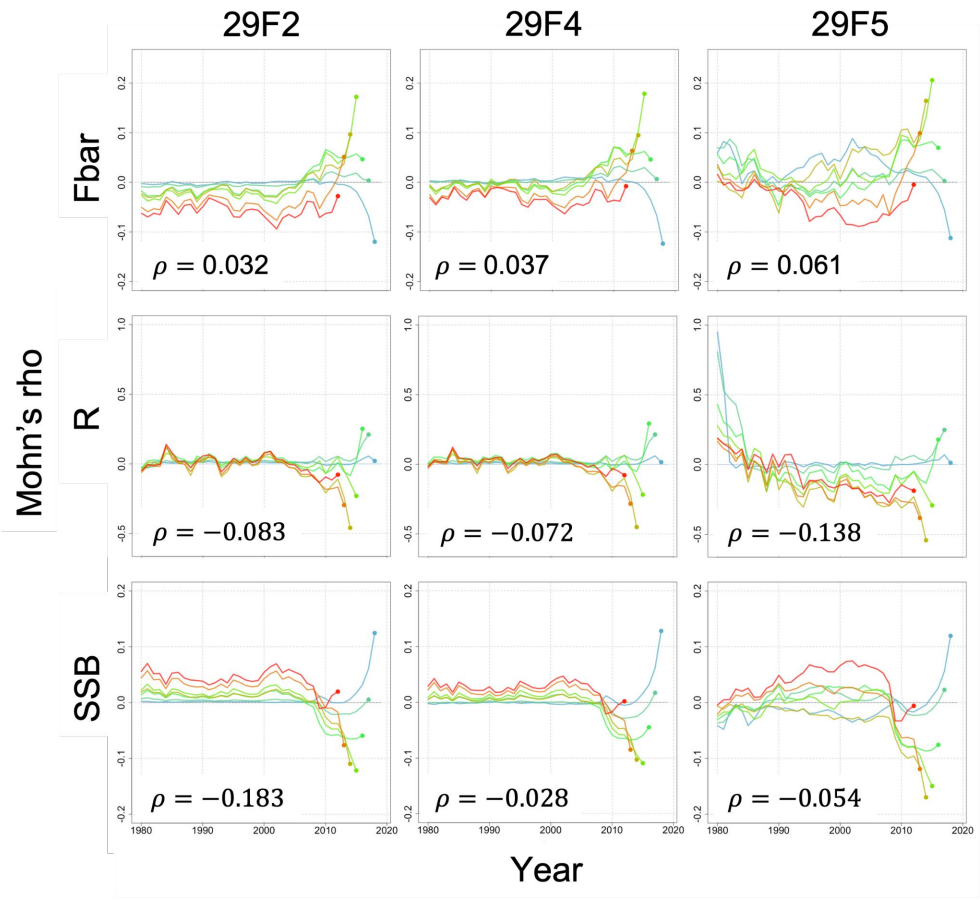

Figure 4.8: Seven year retrospective peels on relative scale for F, R, and SSB for each of the candidate model runs (29F2, 29F4, 29F5).

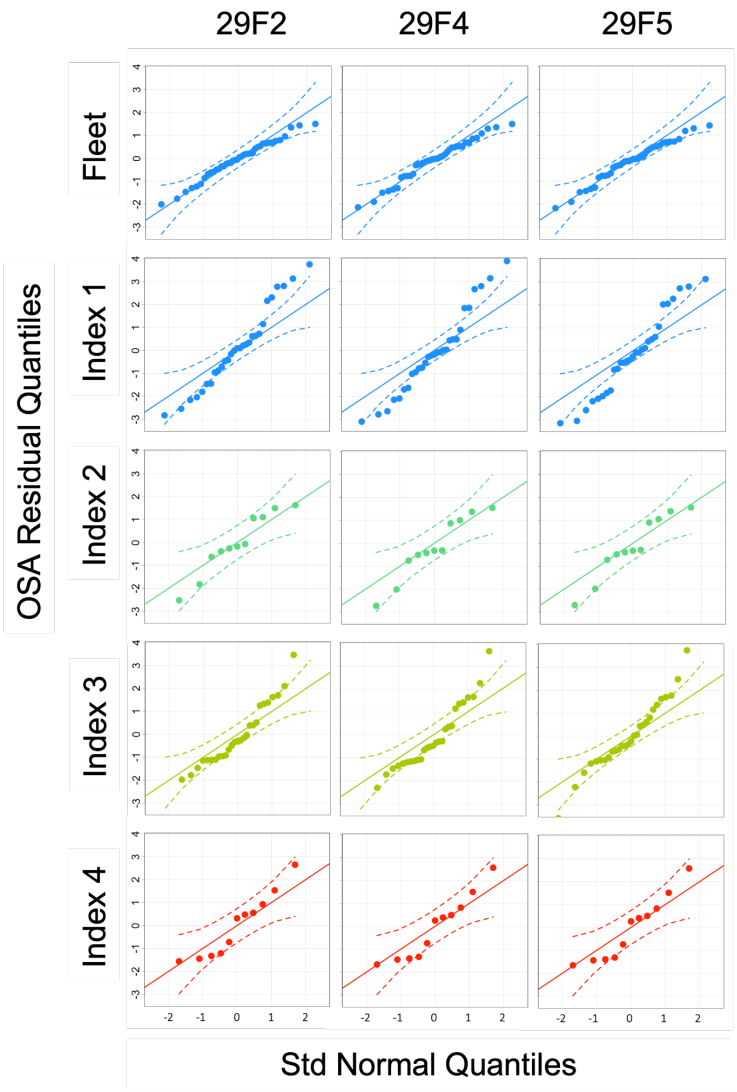

Figure 4.9: QQ plots reflecting the normalcy of OSA residual distributions for fit to aggregate fleet and index data.

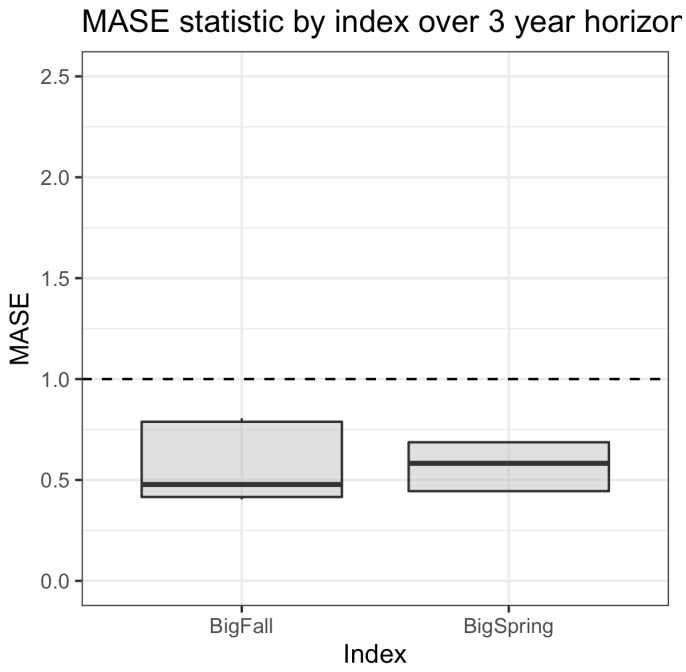

Figure 4.10: Mean absolute scaled error (MASE) calculated over a three year horizon to describe the accuracy with which spring and fall Bigelow indices are predicted by candidate runs. MASE scores less than 1 indicate indices are predicted with more accuracy than a naive approach, with smaller scores reflecting increased accuracy.

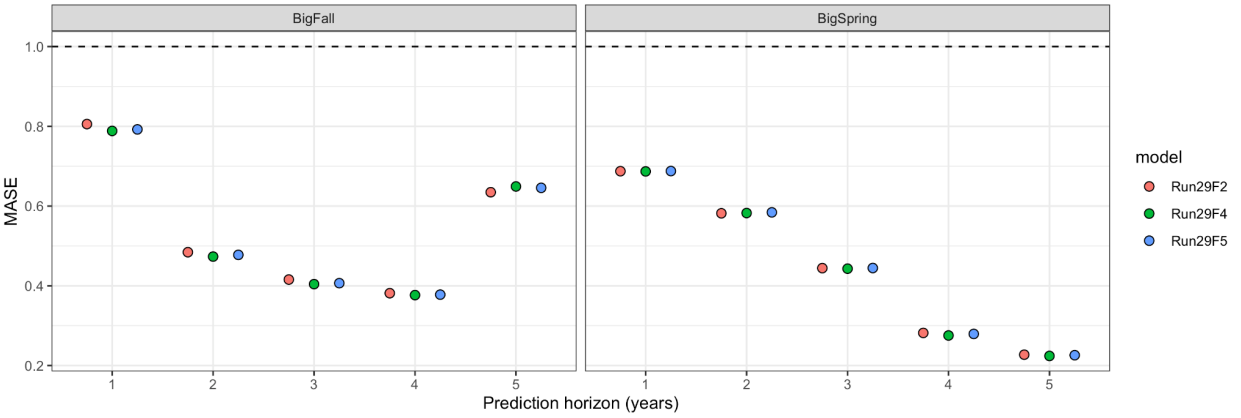

Figure 4.11: Mean absolute scaled error (MASE) for both the spring and fall Bigelow indices over prediction horizons of 1-5 years. MASE scores less than 1 indicate indices are predicted with more accuracy than a naive approach, with smaller scores reflecting increased accuracy.

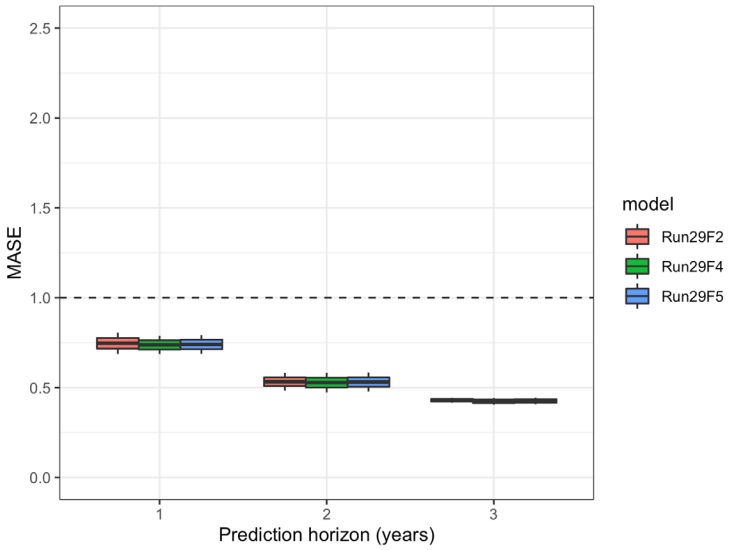

Figure 4.12: Mean absolute scaled error (MASE) calculated over 1-3 year prediction horizons for each candidate model by averaging MASE scores for spring and fall Bigelow indices for each model and horizon. MASE scores less than 1 indicate indices are predicted with more accuracy than a naive approach, with smaller scores reflecting increased accuracy.

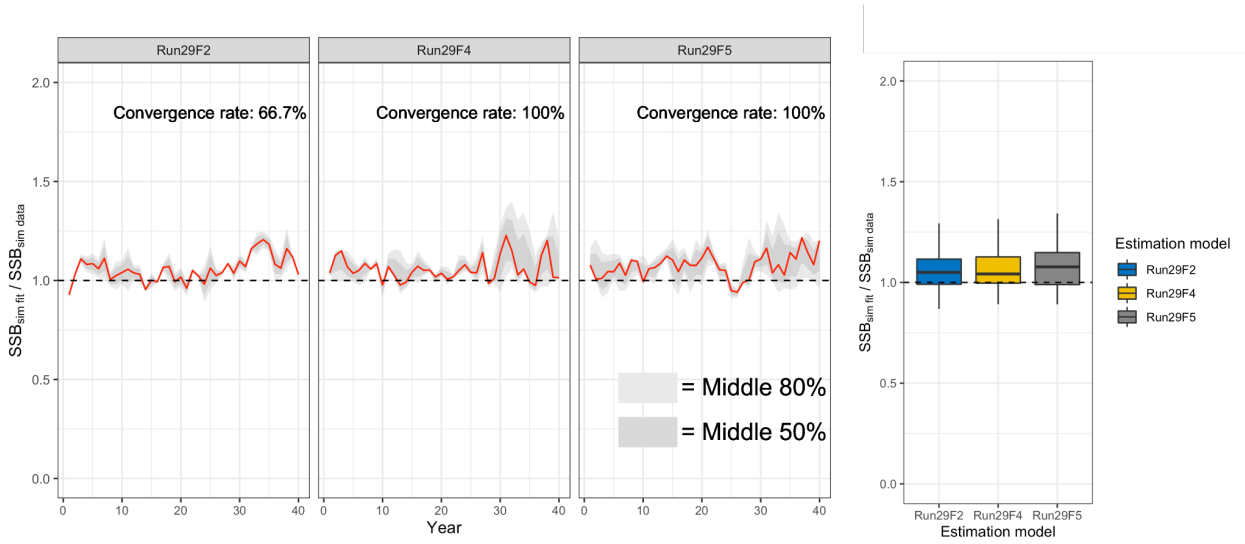

Figure 4.13: Spawning stock biomass (SSB) relative error for candidate model self-tests over time (3 leftmost panels, median in red, middle 50% and 80% of simulations in dark gray and gray respectively). Boxplot of relative error aggregated across all simulations and years for each of the candidate models (rightmost panel).

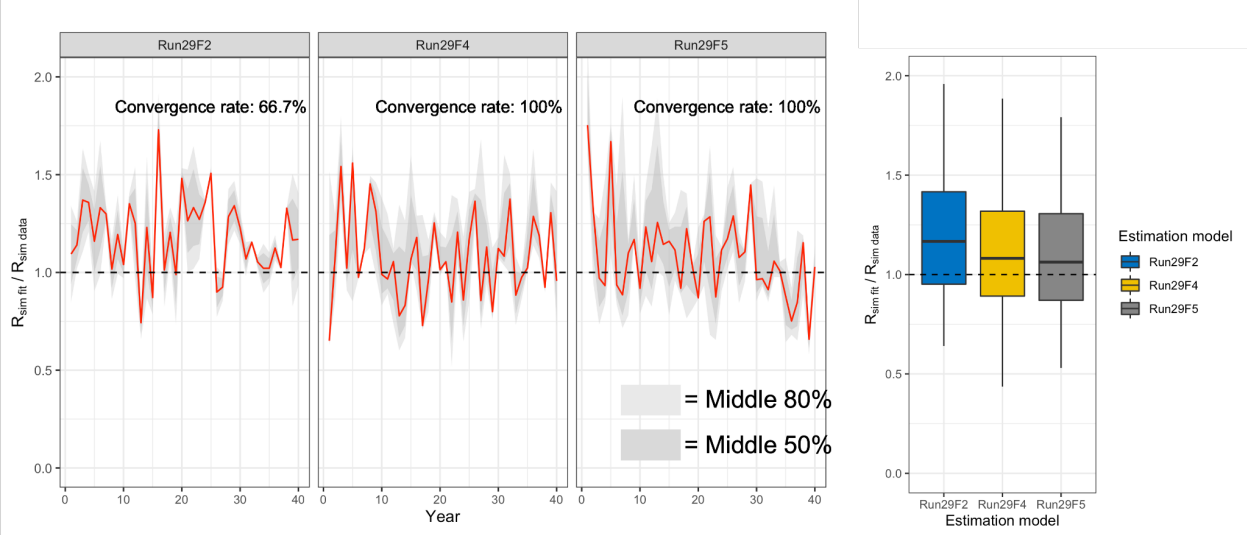

Figure 4.14: Recruitment (R) relative error for candidate model self-tests over time (3 leftmost panels, median in red, middle 50% and 80% of simulations in dark gray and gray respectively). Boxplot of relative error aggregated across all simulations and years for each of the candidate models (rightmost panel).

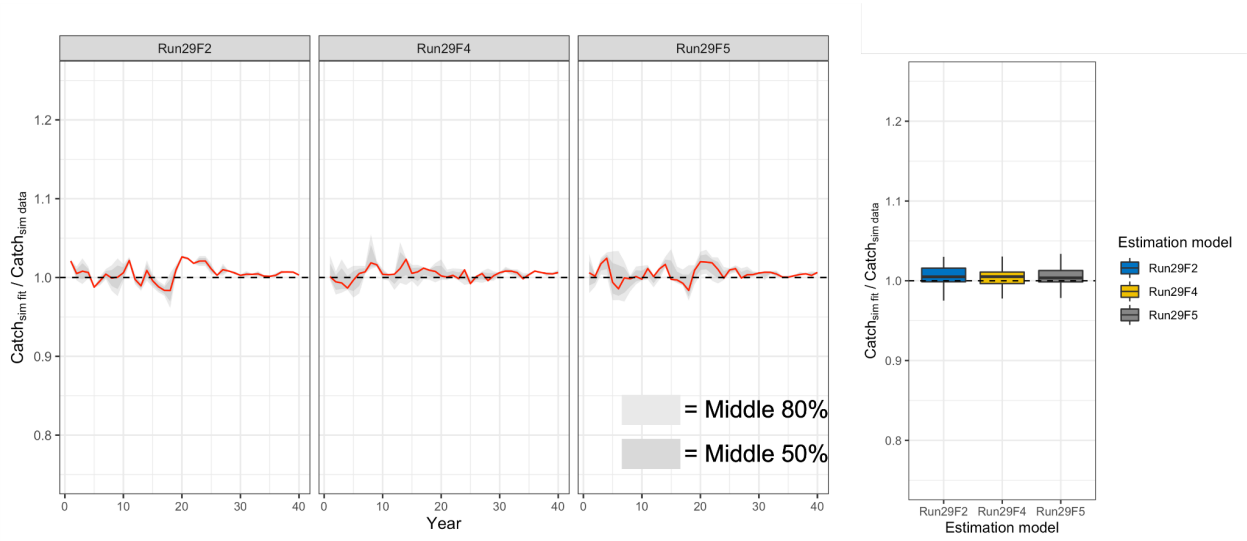

Figure 4.15: Catch relative error for candidate model self-tests over time (3 leftmost panels, median in red, middle 50% and 80% of simulations in dark gray and gray respectively). Boxplot of relative error aggregated across all simulations and years for each of the candidate models (rightmost panel).

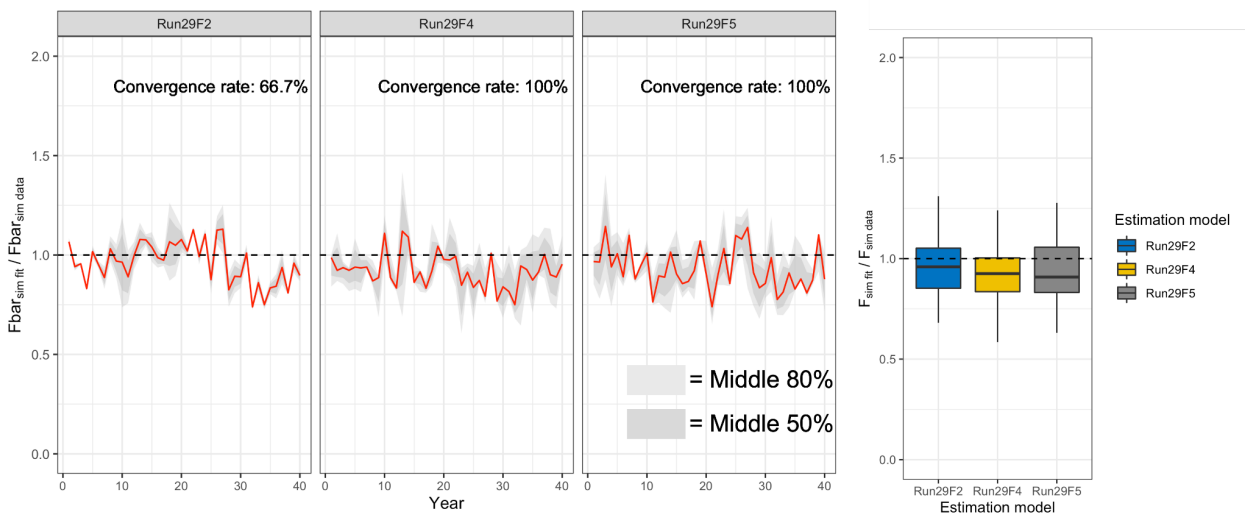

Figure 4.16: Fishing mortality (Fbar) relative error for candidate model self-tests over time (3 leftmost panels, median in red, middle 50% and 80% of simulations in dark gray and gray respectively). Boxplot of relative error aggregated across all simulations and years for each of the candidate models for each of the candidate models (rightmost panel).

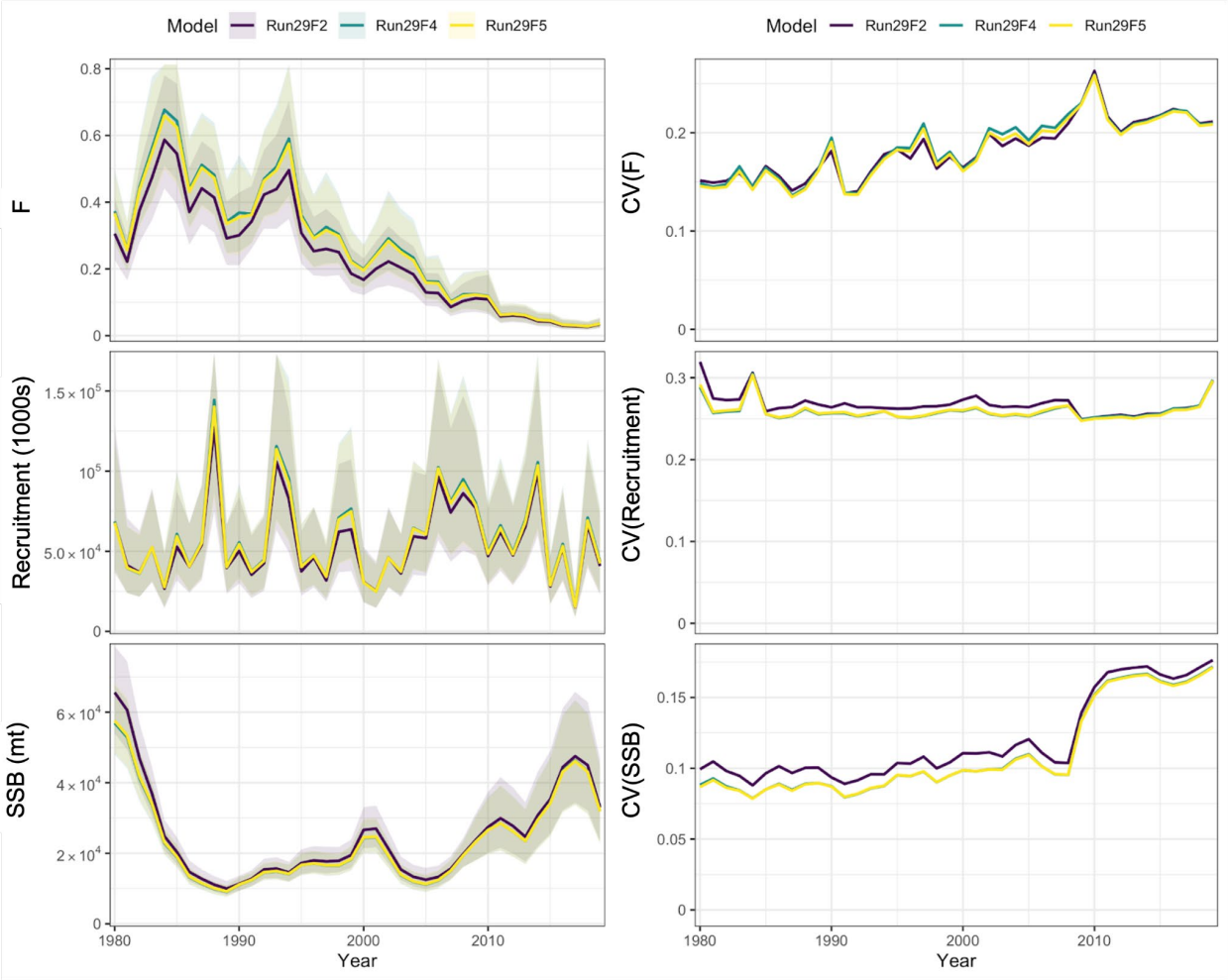

Figure 4.17: Estimates of fishing mortality (F), recruitment, and spawning stock biomass (SSB, left) and CVs around these estimates (right) for candidate runs 29F2, 29F4, and 29F5.