Working Paper 10: Mackerel - Age Structured Assessment Program to Woods Hole Assessment Model

Kiersten Curti, Alex Hansell

Northeast Fisheries Science Center, Woods Hole MA USA

Background and assessment history:

The northwest Atlantic mackerel stock is composed of two spawning contingents: one that spawns in the southern Gulf of St Lawrence primarily during July and another that spawns primarily in southern New England in May/June (Sette 1950). Atlantic mackerel migrate seasonally and during winter months, the two contingents mix on the Northeast U.S. shelf and are available to the U.S. fishery.

Northwest Atlantic mackerel was historically assessed in the U.S. using Virtual Population Analysis (VPA). Assessments completed during the 1990s generally indicated low fishing mortality rates and high spawning stock biomasses, however, there were concerns over imprecise estimates, highly correlated residuals, and the possibility of a non-linear relationship between the NEFSC trawl survey and stock abundance (e.g., NEFSC 1996a). Additionally, a VPA completed for the 2000 assessment was not used to determine stock status due to a lack of convergence, survey variability and a strong retrospective pattern in spawning stock biomass (NEFSC 2000). During the mid-2000s, a statistical catch-at-age (ASAP) model was accepted that again indicated low levels of fishing mortality and high SSB estimates, however, the reviewers noted a lack of larger, older fish in both fishery and survey catches, as well as the presence of significant retrospective patterns in SSB, fishing mortality and recruitment (NEFSC 2006). In the late 2000s, a transboundary assessment was completed using a VPA, but this model was not accepted for use in management due to poor diagnostics, model instability and significant retrospective patterns (Deroba et al. 2010). For this assessment, the reviewers noted that much of the problem stemmed from conflicts among datasets, in particular between the NEFSC spring bottom trawl survey and fishery catch-at-age time series. In reviewing the results of the 2006 assessment, this assessment also concluded that due to a significant retrospective pattern, the reference points from the previous assessment were also considered to be inappropriate.

Since the 2017 benchmark assessment (NEFSC 2018), northwest Atlantic mackerel has been assessed using ASAP. The ASAP model accepted during this benchmark assessment incorporated a new rangewide egg index that helped to resolve the disparate trends between the fishery and survey time series, was robust to model assumptions, and only exhibited a minimal retrospective pattern (SSB Mohn's rho of 0.16). However, this model exhibited a poor fit to the NEFSC trawl survey Albatross index, potentially indicating the presence of process errors that were not fully captured in the assessment model. When the model was updated for the 2021 and 2023 Management Track assessments, the retrospective patterns increased in magnitude (NEFSC 2021, NEFSC 2023). Additionally, these management track assessments also indicated that the increase in spawning stock biomass projected during the previous assessments were not realized. In particular, rebuilding projections resulting from the 2017 benchmark assessment (NEFSC 2018) projected a spawning stock biomass of approximately 162,796 mt in 2019; however, the 2021 management track assessment estimated a terminal year SSB of 42,862 mt for 2019, which was then updated to 23,505 mt during this management track. Likewise, the 2021 management track assessment projected a spawning stock biomass of approximately 70,768 mt in 2022; however, this assessment estimated a terminal year SSB of 19,017 mt for 2022. During the most recent (2023) management track assessment, the review panel recommended the development of a state-space model to better deal with process error and changing ecosystems. Here, we explore fitting the data inputs from the 2023 ASAP assessment to the Woods Hole Assessment Model (WHAM) and propose a base WHAM run that can be further enhanced for the 2025 management track assessment.

Data inputs and ASAP parameterization:

A full description of the data inputs to the Atlantic mackerel assessment model is provided in NEFSC (2018). The most recent ASAP model incorporated ages 1-10+, fishery and survey data from 1968-2022, and three index time series (Figure 1-3). Combined Canadian and U.S. fishery catch (commercial and recreational) were modeled as one fishing fleet with constant selectivity over time. Fishery selectivity was assumed to be flat-topped with age-specific selectivity parameters fixed at one for ages 6+. The primary survey dataset is an index of spawning stock biomass that is developed from a dedicated egg survey in Canadian waters and the May/June MARMAP (1977-1988) and EcoMon (1992-2022) ecosystem surveys in U.S. waters. Additionally, the NEFSC spring bottom trawl survey is incorporated into the model and is treated as two distinct time series for the Albatross (1968-2008) and Bigelow (2009-2022) years. For these time series, only ages 3+ are used to develop the index because recent otolith microchemistry work demonstrated that these ages were most representative of the unit stock and not just the local spawning contingent (Redding et al. 2020). Trawl survey selectivity is fixed at one for age-3 with ages 4+ estimated parameters. Fishermen have suggested that due to the NEFSC trawl survey's tow speed and length, the trawl survey is likely better able to catch smaller fish than larger individuals (NEFSC 2018). Natural mortality was assumed to be 0.2 and constant over both time and age. Annual maturity ogives developed from Canadian samples representing the northern spawning contingent were used. This assumption is consistent with trends in the egg index, which indicate that the majority of the spawning stock is composed of individuals from the northern contingent.

WHAM model configurations:

In this study, the same data inputs were used as the 2023 management track assessment. No changes were explored regarding natural mortality or environmental covariates. For exploring different possible age composition likelihoods, input effective sample sizes were increased to be 200 for all data sources and years. High input effective sample sizes were needed for the self-weighting age composition distributions.

Model exploration was completed in the following order: 1) age distributions; 2) recruitment assumptions; 3) time-varying selectivity; 4) full state-space models. WHAM allows for 10 different age-composition distributions and all were explored for mackerel. Different recruitment assumptions were also explored, including: random walk, beverton-holt stock recruitment, and mean recruitment where deviations from the mean were independent and identically distributed (iid) or correlated by year with an ar1 process. Time-varying selectivity in the fishing fleet was explored. When exploring time-varying selectivity age-6 was fixed at one and all other ages were freely estimated. A series of state-space models were also explored with random effects on numbers-at-age.

The same diagnostics used during the most recent Atlantic Cod Research Track were used to inform model selection:

1. Convergence: This diagnostic was used to determine if models produced a stable solution. Both first and second order convergence had to occur. First order convergence was determined by the maximum gradient of the negative log-likelihood function being less than e^-4. Second order convergence required the Hessian matrix being invertible.
2. Residuals: Goodness of fit was determined by examining model fit to the different data inputs. One-step ahead (OSA) residuals were used as the primary diagnostic because they help to account for correlations in the data that can be propagated into Pearson residuals (Trijoulet et al 2023). OSA residuals should be normally distributed.
3. Akaike's Information Criteria (AIC): AIC was used to help determine parsimony and compare models with comparable likelihoods. Smaller AIC scores are better, with an improvement in model being determined if the difference in AIC between two models was greater than two.
4. Retrospective patterns: Retrospective error was explored by comparing seven-year Mohn's rho value for spawning stock biomass and fishing morality.
5. Prediction skill: Prediction skill was estimated by using mean absolute scaled error (MASE). A 10-year peel was used, with prediction error calculated over one- to five-year time periods. First-order model convergence was calculated for each peel. A MASE below one indicates predictive skill that is better than a naïve approach of predicting future observations.
6. Estimation performance: Self-tests were used to explore estimation performance. The final candidate model was used to generate 100 datasets with parameters fixed at their estimated values. Simulations were then refit the model to these datasets to investigate relative error in F, SSB, R and convergence rates.

Results:

All ASAP material from the last Management Track assessment as well as current WHAM runs are available on GitHub: https://github.com/kcurti/Mack.WHAM.Development. See "run_notes.doc" on GitHub for a detailed description of each WHAM run.

A WHAM run configured the same way as ASAP successfully converged and produced similar trends as ASAP (Run1; Figure 4). Selectivity estimates were similar to original ASAP runs for the fishing fleet and the three indices (Figure 5-6). Runs exploring the different age-compositions supported using a logistic normal age distribution that used an ar1 process to estimate zero values; the runs using this distribution resulted in similar temporal trends but had improved residual patterns and similar Mohn's rho values to other age composition likelihoods (Table 1; Figure 7). Thus, the logistic normal age distribution with an ar1 process to estimate zero values was used in subsequent model runs.

Incorporating random effects into recruitment alone did not substantially improve model diagnostics so they were not included (Table 2; Figure 8; Runs 4-6 on Github). Models fit with a Beverton-holt stock recruit relationship had similar diagnostics as models that assumed mean recruitment; however, the Beverton-holt stock recruit relationship did not fit well to observed larger SSB values (Run 6). Models with random effects on fleet selectivity converged but did not significantly improve diagnostics (Table 3; Figure 9-10; Runs 8-10 on Github). Additionally, there were concerns that some of the estimated trends in time-varying selectivity did not reflect observed cohorts.

Full state-space models converged and including 2dar1 random effects on numbers-at-age led to improved Mohn's rho values (Table 4; Figure 11-12; Run 12-15 on Github). Thus, the selected model had similar inputs to the original ASAP model; however, it had logistic-normal distributions with an ar1 process on all age composition likelihoods and random effects using a 2dar1 process on numbers-at-age.

The selected WHAM run (Run 13) had similar fits as ASAP to the aggregate catch and survey indices (Figure 14-17). The proposed WHAM run had decent fits to the age composition data (Figure 18-20) and had better retrospective patterns than ASAP (Table 5; Figure 21). The MASE scores were comparable between the selected WHAM run and other runs (Table 2-4; Figure 22). MASE scores were given less weight during model selection because of their limited utility for state-space models (Chengxue et al. Working Paper 3). The self-test on the selected model produced mean bias below 10% for F, SSB and R (Table 6; Figure 23). Reference points from WHAM were similar to those from ASAP but smaller in magnitude (Table 7); however, stock status was robust to these differences (Figure 24). The proposed WHAM run had similar trends as ASAP; however, it estimated larger uncertainty in SSB, F and consequently, stock status (Figure 25). This increased uncertainty is presumably due to the large process error estimates for recruitment and survival (2dar1 random effects on NAA); however, future work should further evaluate the trade-offs between model diagnostics and uncertainty estimates.

Short-term projections were completed at $F_{msy\ proxy}$. These WHAM projections exhibited the similar temporal trends as those from the ASAP model, but had much larger confidence intervals associated with the projected estimates (Table 8, Figure 26).

Discussion:

Overall, the WHAM run produced similar trends as the previous ASAP model. There are several advantages from moving to WHAM: 1) the logistic-normal age compositions are self weighting; 2) including 2dar1 random effects on numbers-at-age reduces retrospective patterns; 3) the 2dar1 random effects on numbers-at-age also helps to account for recruitment correlations. However, the proposed WHAM run also has increased uncertainty around key parameter estimates. More work is needed to further determine if these uncertainties are realistic or an artifact of model specification. The base WHAM model that was configured like ASAP (fixed effects only) had similar trends to ASAP and similar estimates of uncertainty. At a minimum, moving over to WHAM and configuring it the same as ASAP would allow for future model development.

Transitioning from ASAP to WHAM will also allow for future work in the parameterization of recruitment, selectivity and catchability that can continue to improve the assessment.

For example, with the exception of the 2015 and 2021 year classes, recruitment has been below the time-series median since 2009. A change-point analysis of recruits-per-spawner showed greater support that low recruitment was due to a depressed SSB than environmentally driven but a clear stock-recruitment relationship was not apparent (NEFSC 2023). Moreover, the poor-performance of short-term projections has been partially due to an overestimation of projected recruitment. Accordingly, the parameterization of recruitment as a mean plus deviations that are correlated over both age and year may enhance the reliability of short-term projections but future recruitment parameterizations like a random walk could also be explored.

Additionally, there are a priori reasons why the catchability of the NEFSC bottom trawl survey may vary over time. The NEFSC spring bottom trawl survey does not cover the full geographic range of the northwest Atlantic mackerel stock and annual variability in seasonal migration patterns, particularly the movement of the northern spawning contingent into U.S. waters, could impact the proportion of the stock available to the trawl survey. These potential changes in availability could explain the poor fit to the Albatross time series in the ASAP model. Time-varying catchability in the survey indices was explored during model development in this Research Track; however, models struggled to converge, especially when additional random effects were included. Consequently, time-varying catchability (ar1_y or iid) in the NEFSC bottom trawl surveys is an avenue for future exploration.

Likewise, there are also a priori reasons why fishery selectivity may vary over time. The primary gears in the U.S. fishery are midwater and bottom trawls where Canadian landings are primarily from gillnets, seines and fixed gear. Because U.S. and Canadian catches are combined and modeled as one fishing fleet, varying contributions of the fisheries to the total annual catch could impact the apparent selectivity. Random effects in fishery selectivity were estimated in runs 8-11; however, these runs produced unrealistically low age-specific selectivity estimates in the early 2000s when the strong 1999 cohort was known to move through the fishery. Consequently, parameterization of time-varying selectivity should continue to be explored in future WHAM runs.

References:

Chengxue Li, Jonathan J. Deroba, Timothy J. Miller, Christopher M. Legault, Charles T. Perretti (Working Paper 3). An evaluation of common stock assessment diagnostic tools for choosing among state-space 2 models with multiple random effects processes.

Deroba, J.J., G. Shepherd, F. Gregoire, J. Nieland, J. Link. 2010. Stock Assessment of Atlantic Mackerel in the Northwest Atlantic - 2009. TRAC Reference Document 2010/01.

NEFSC (Northeast Fisheries Science Center). 1996a. Report of the 20th Northeast Regional Stock Assessment Workshop (20th SAW): Stock Assessment Review Committee (SARC) Consensus Summary of Assessments. NMFS NEFSC Ref. Doc. No. 95-18. 219 p.

NEFSC (Northeast Fisheries Science Center). 2000. 30th Northeast Regional Stock Assessment Workshop (30th SAW): Public Review Workshop. NMFS NEFSC Ref. Doc. No. 00-04. 58 p.

NEFSC (Northeast Fisheries Science Center). 2006. 42nd Northeast Regional Stock Assessment Workshop (42nd SAW) Stock Assessment Report Part A: Silver Hake, Atlantic Mackerel, and Northern Shortfin Squid. NMFS NEFSC Ref. Doc. No. 06-09a. 290 p.

NEFSC (Northeast Fisheries Science Center). 2018. 64th Northeast Regional Stock Assessment Workshop (64th SAW) Assessment Report. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 18-04; 529 p.

NEFSC (Northeast Fisheries Science Center). 2021. Management Track Assessment: June 2021. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 22-10; 79 p. https://doi.org/10.25923/4m8f-2g46

NEFSC (Northeast Fisheries Science Center). 2023. Management Track Assessments Spring 2023. US Dept Commer, Northeast Fish Sci Cent Tech Memo. 308; 69p.+x. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Redding, S. G., L. W. Cooper, M. Castonguay, C. Wiernicki, and D. H. Secor. 2020. Northwest Atlantic mackerel population structure evaluated using otolith d18O composition. ICES Journal of Marine Science 77: 2582-2589. https://doi.org/10.1093/icesjms/fsaa117

Sette, O.E. 1950. Biology of the Atlantic Mackerel (*Scomber scombrus*) of North America, Part II: Migrations and Habits. Fishery Bulletin 49, Volume 51: 251-358.

Trijoulet, V., Albertsen, C. M., Kristensen, K., Legault, C. M., Miller, T. J., & Nielsen, A. (2023). Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. Fisheries Research, 257, 106487.

Tables

Table 1: Age composition likelihood assumptions explored during WHAM model development.

age_comp	converged	hessian	NLL	rho_R	rho_SSB	rho_Fbar
multinomial	Yes	No	2305.805	0.1608	0.3633	-0.228
dir-mult	Yes	Yes	1769.636	0.5752	0.8363	-0.4225
dirichlet-miss0	Yes	Yes	-1488.64	0.9747	1.1804	-0.4947
dirichlet-pool0	Yes	Yes	-1481.17	0.9476	1.1474	-0.4894
logistic-normal-miss0	Yes	Yes	-1344.85	0.7369	0.7068	-0.2969
logistic-normal-ar1-miss0	Yes	Yes	-1578.05	0.8776	0.8802	-0.4192
logistic-normal-pool0	Yes	Yes	-1319.17	0.6501	0.6585	-0.2722
mvtweedie	Yes	Yes	-1319.17	0.1809	0.5006	-0.2307

Table 2: Recruitment assumptions explored during WHAM model development.

Treatment	dAIC	AIC	rho_R	rho_SSB	rho_Fbar	MASE egg survey	MASE Bigelow
None	0	-3212.3	0.8776	0.8802	-0.4192	0.98	0.53
ar1	159.1	-3053.2	1.0685	1.1322	-0.4988	0.852	0.491
iid	168.6	-3043.7	1.2238	1.3098	-0.5429	1.14	0.648
Beverton holt	190.8	-3021.5	0.7216	0.9133	-0.4501	1.39	0.693

Table 3: Different time varying selectivity assumptions explored during WHAM development.

Treatment	dAIC	AIC	rho_R	rho_SSB	rho_Fbar	Mase egg index	MASE Bigelow
None	0	-3212.3	0.9922	1.0122	-0.4596	0.98	0.53
2dar-1	327.6	-2884.7	0.8204	0.9282	-0.3835	1.02	0.583
iid	330.4	-2881.9	0.6096	0.7036	-0.3622	0.964	0.562
ar1_y	342.2	-2870.1	0.8776	0.8802	-0.4192	0.955	0.556

Table 4: Different state-space models explored during WHAM development.

Treatment	dAIC	AIC	rho_R	rho_SSB	rho_Fbar	MASE egg survey	MASE Bigelow
none	0	-3212.3	0.9922	1.0122	-0.4596	0.98	0.53
ar1_a	152	-3060.3	1.0909	1.2315	-0.5351	0.91	0.47
ar1_y	190.8	-3021.5	0.7216	0.9133	-0.4501	0.906	0.477
iid	208.9	-3003.4	0.969	1.1679	-0.5184	1.01	0.50
2dar1	214.2	-2998.1	0.2514	0.2808	-0.1624	1.49	0.832

Table 5: A comparison of Mohn's rho values between the accepted ASAP and proposed WHAM models.

VAR	ASAP	WHAM
F	-0.306	0.162
SSB	0.485	0.281
R	0.132	0.251

Table 6: Mean percent bias from the self-test of the proposed WHAM model.

VAR	Mean Percent Bias
F	-0.71
R	7.25
SSB	5.67

Table 7: A comparison of biological reference points estimated from the accepted ASAP and proposed WHAM models.

BRP	**ASAP**	**WHAM**
Fmsy proxy	0.21	0.16
SSBmsy proxy (mt)	154,107 (86,490 – 332,677)	128,150 (49,587 – 331,184)
MSY proxy (mt)	30,460 (17,321 – 63,448)	23,724 (9,123 – 61,693)

Table 8: Comparison of short-term projection estimates at $F_{msy\ proxy}$ from the accepted ASAP and proposed WHAM models.

Year	Parameter	WHAM	lo	hi	ASAP	lo	hi
2023	SSB	56696	14159	227025	43721	17764	100647
2024	SSB	79974	12770	500858	67524	32341	143307
2025	SSB	87139	7424	1022820	77038	43184	156933
2026	SSB	83319	3395	2044805	81551	47719	176672
2023	F	0.081	0.018	0.37	0.15	0.07	0.34
2024	F	0.163	0.147	0.181	0.21	0.21	0.21
2025	F	0.163	0.147	0.181	0.21	0.21	0.21
2026	F	0.163	0.147	0.181	0.21	0.21	0.21
2023	Catch	5953	5953	5953	5953	5953	5953
2024	Catch	15429	2162	110092	12618	6127	27331
2025	Catch	16285	1234	214936	14671	8301	30242
2026	Catch	15422	571	416762	15694	9268	33414
2023	Rect	131200	7204	2389543	91342	28584	294287
2024	Rect	115979	2853	4714488	107496	29362	405817
2025	Rect	102721	1329	7940735	111252	29651	470921
2026	Rect	91150	683	12168829	116499	29937	524124

Figures

A) B)

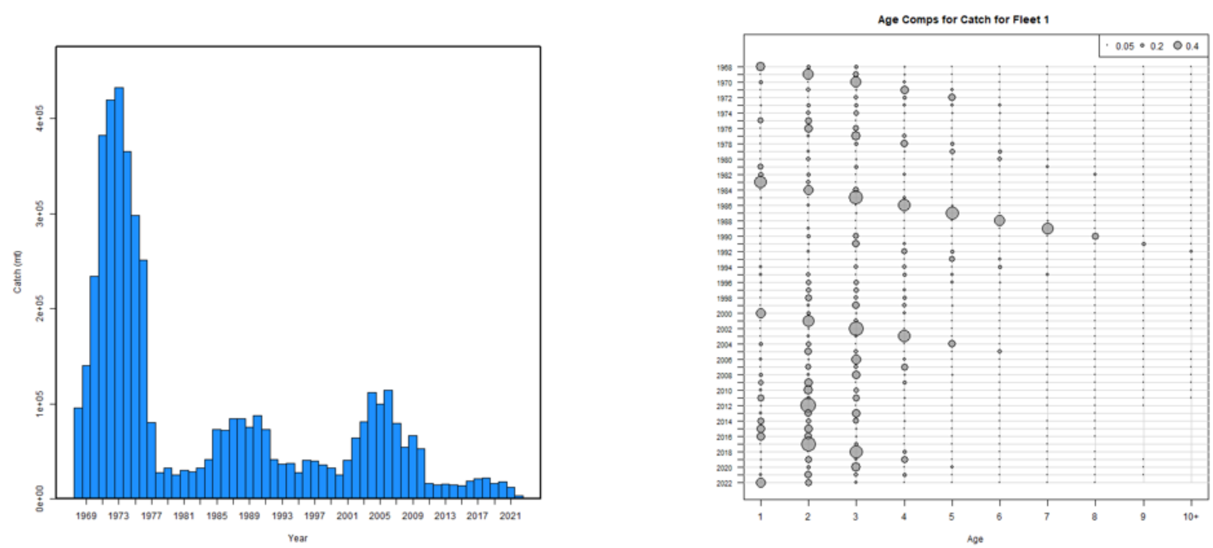

Figure 1: Time series of aggregate fishery catches (a) and corresponding age proportions (b) that were incorporated into the ASAP and WHAM models.

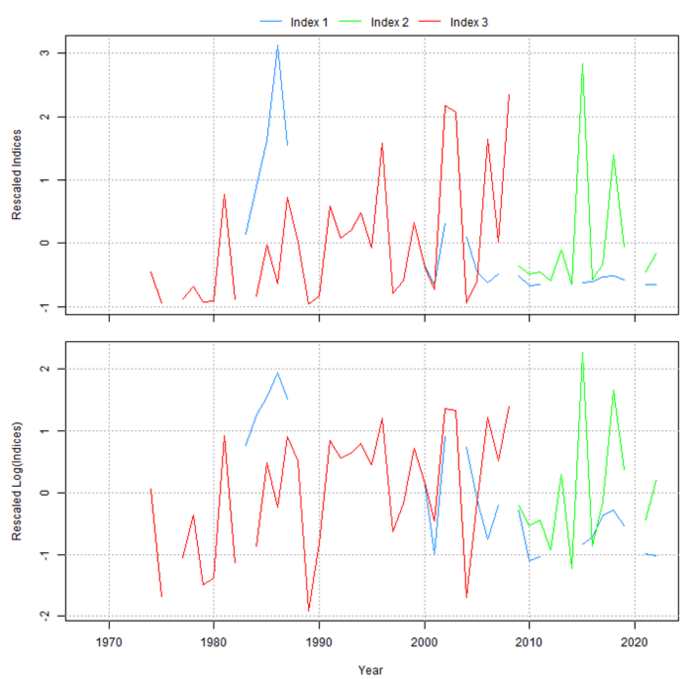

Figure 2: Index time series that were incorporated into the accepted ASAP and proposed WHAM models. (index 1 = egg survey; index 2 = NEFSC Albatross; index3 = NEFSC Bigelow)

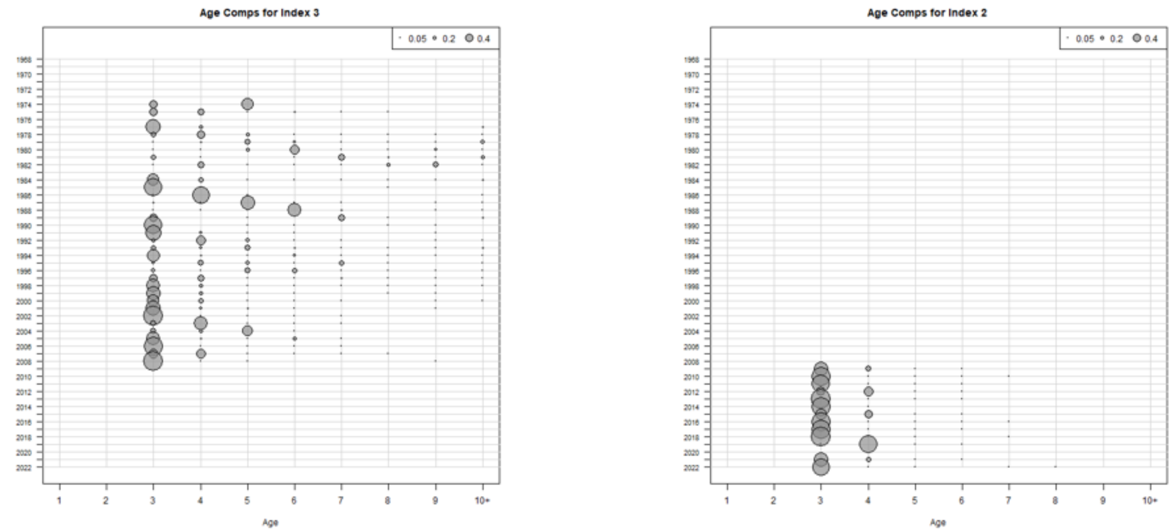

Figure 3: Age compositions for the NEFSC spring bottom trawl surveys (2 = Bigelow; 3 = Albatross).

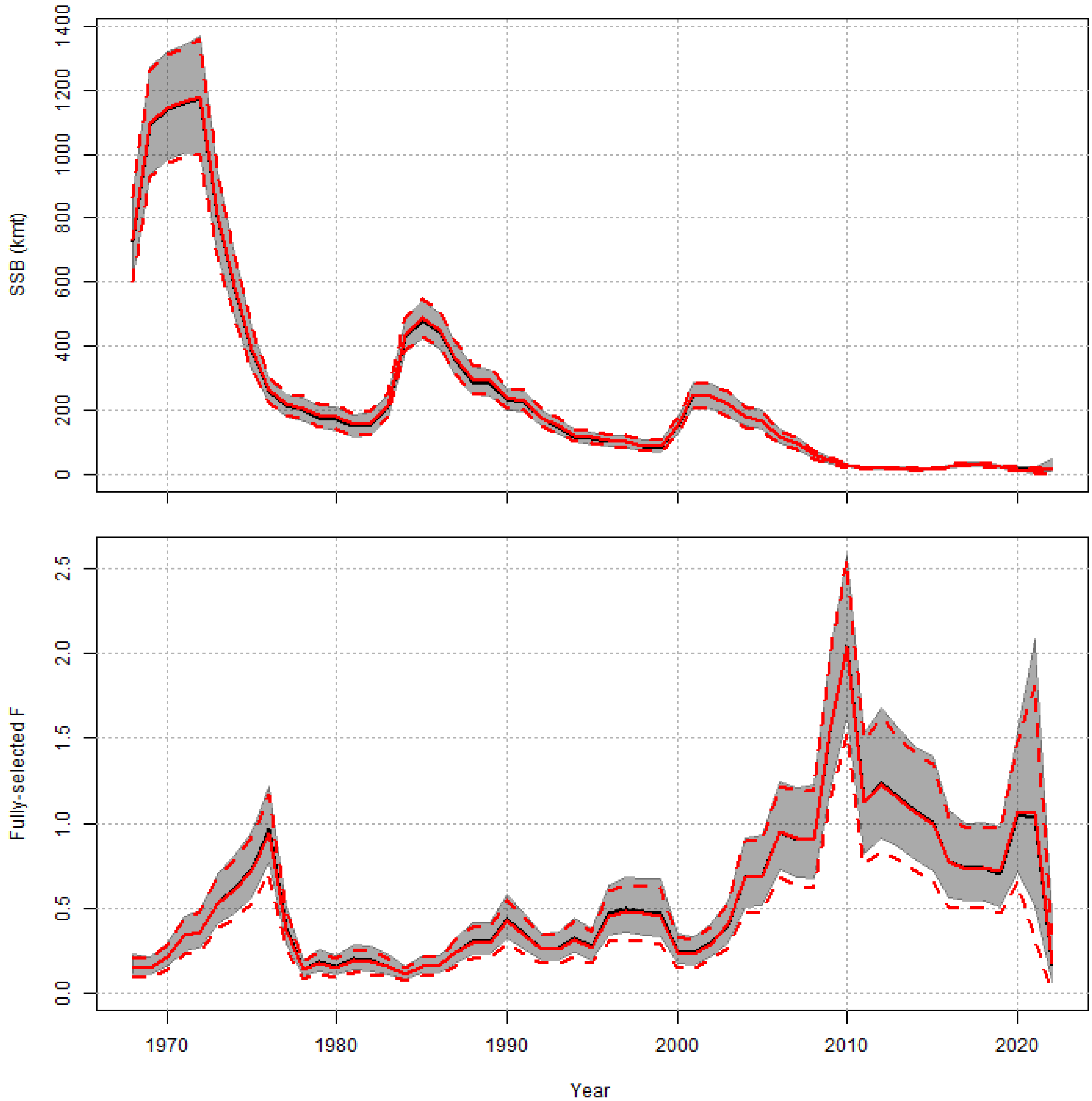

Figure 4: Comparison of spawning stock biomass and fishing mortality time series from the accepted ASAP model (red) used in the last management track and the WHAM run (black) using the same settings.

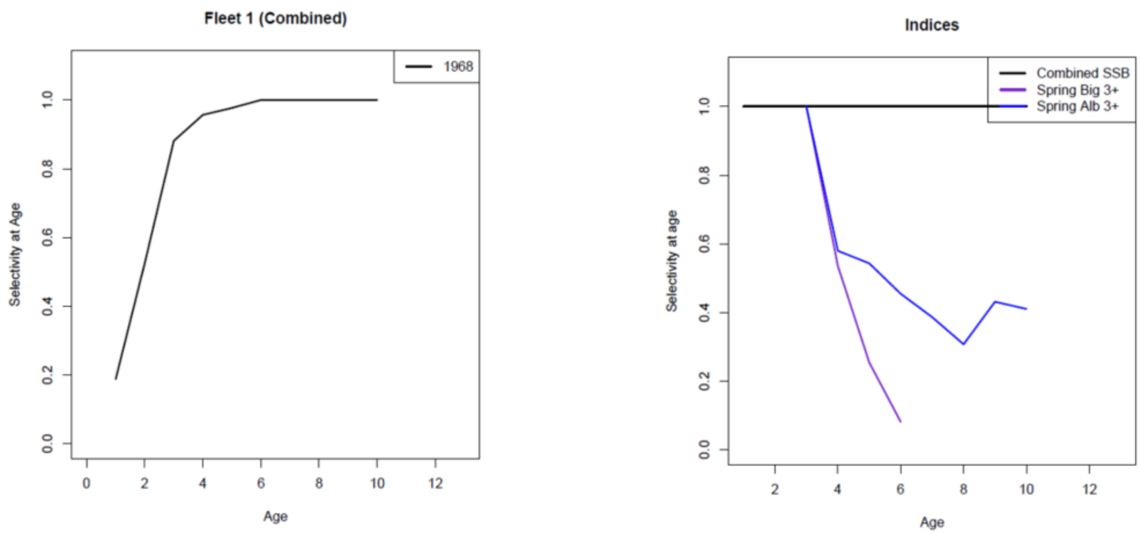

Figure 5: Fishery and survey selectivity estimates from the accepted ASAP model.

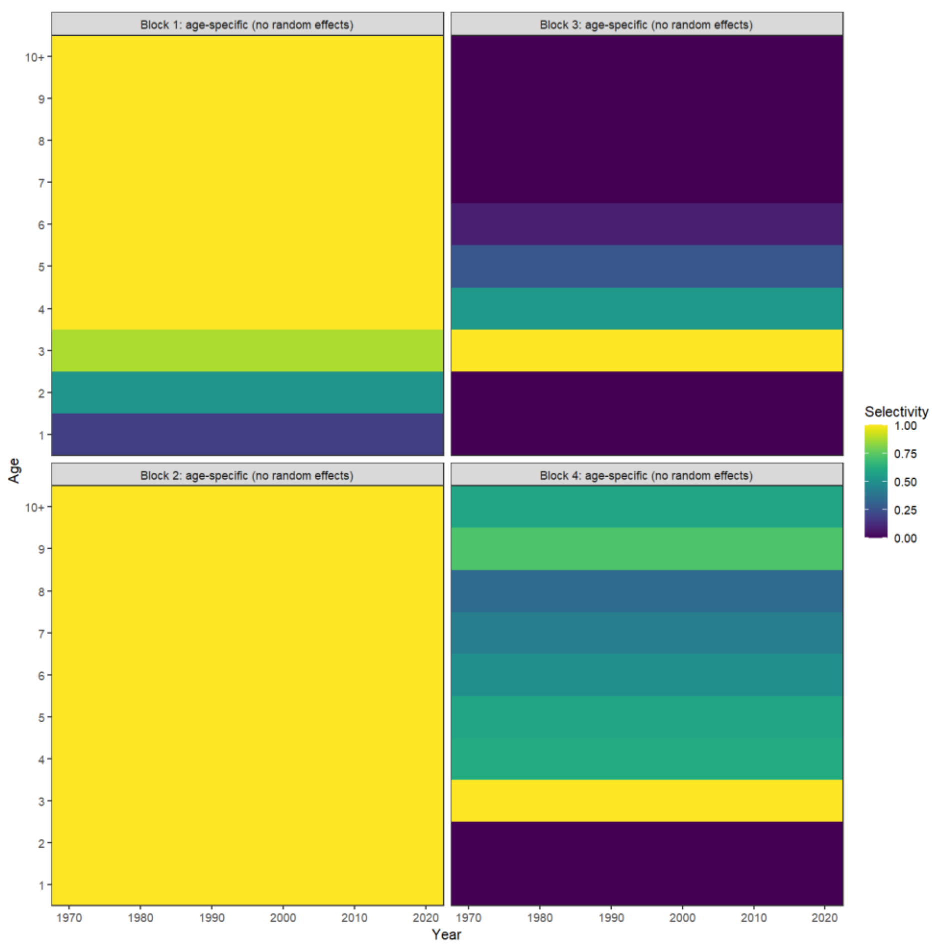

Figure 6: Selectivity estimates for the fishing fleet (block 1), the combined egg index (block 2), NEFSC bigelow (block 3), and NEFSC albatross (block 4) from the proposed WHAM model.

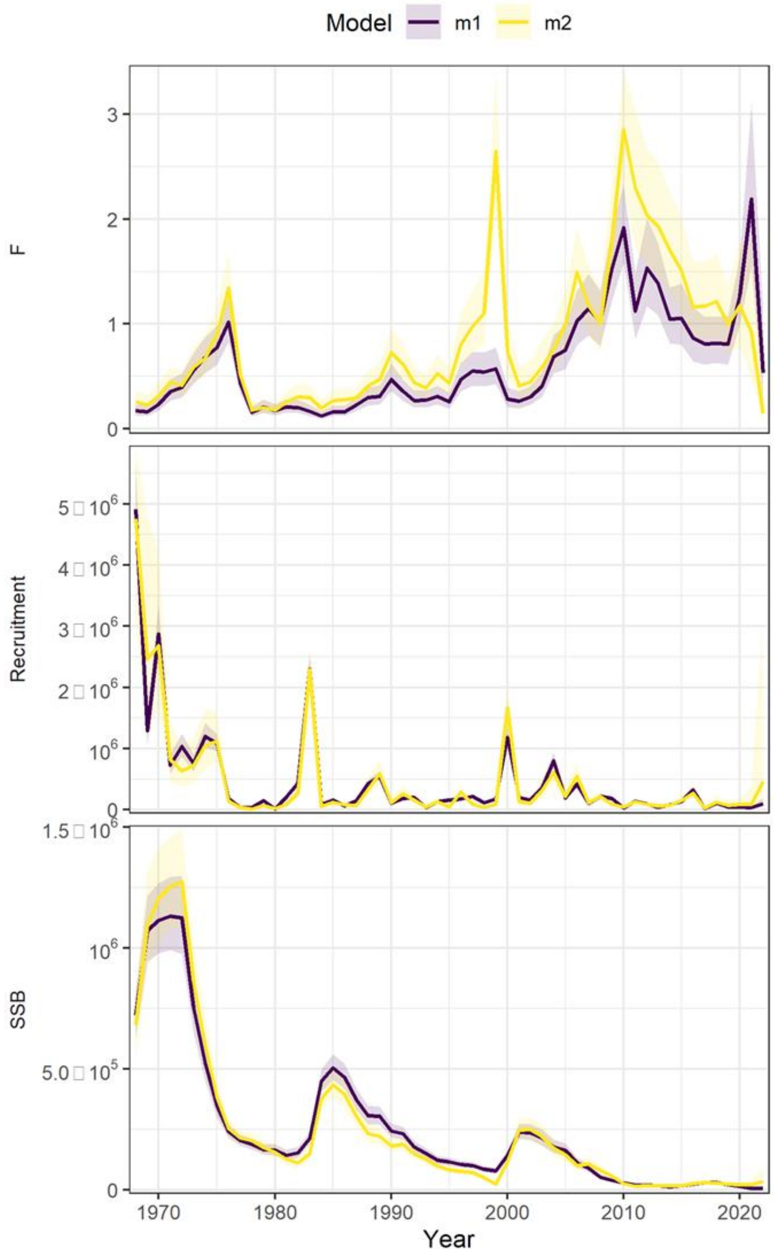

Figure 7: Comparison of key parameter estimates with models fit to multinomial (m1) and logistic normal age (m2) composition likelihood distributions.

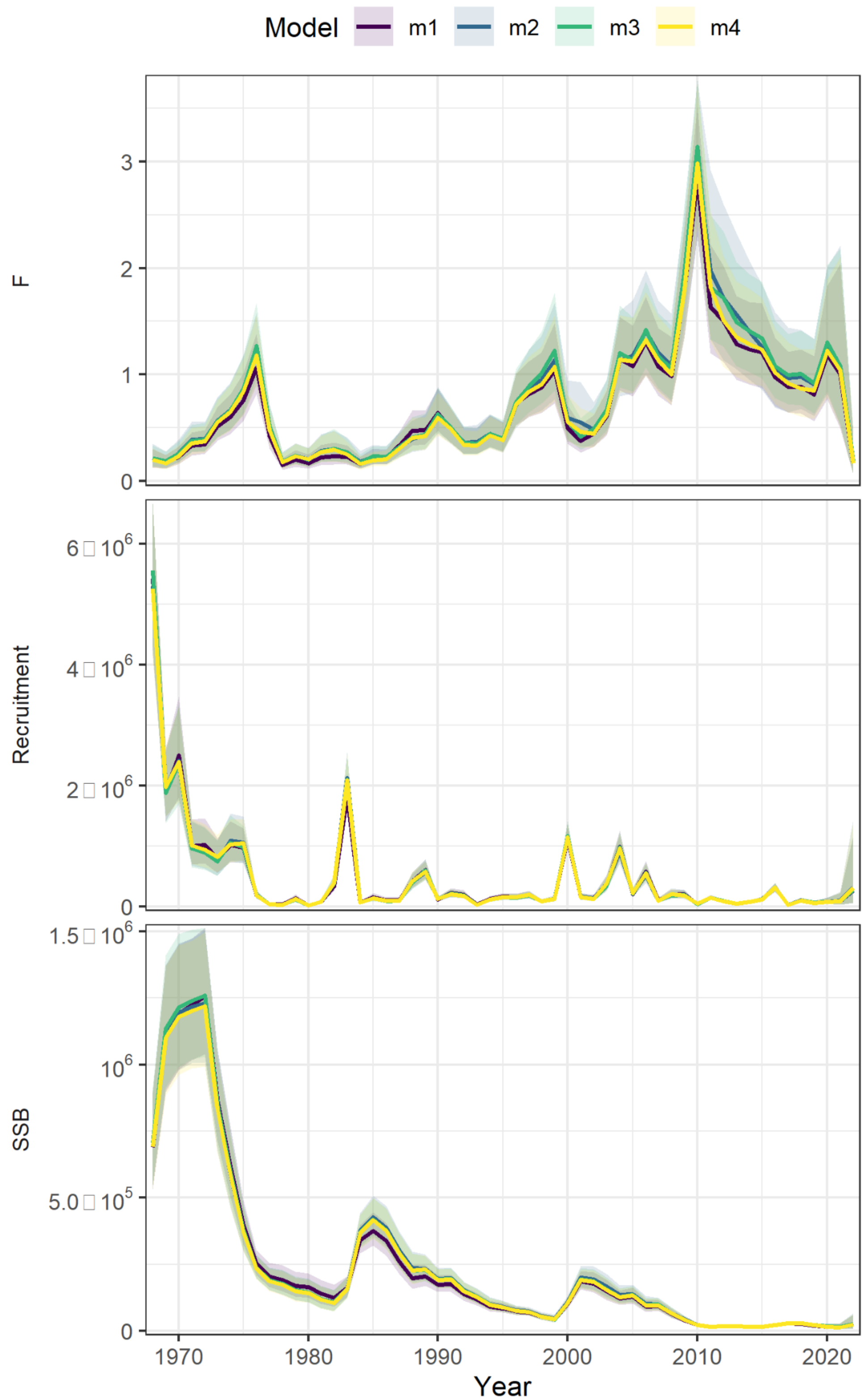

Figure 8: Comparison of runs with different time-varying recruitment assumptions (m1 = none; m2 = ar1_y; m3 = iid; m4 = Beverton holt).

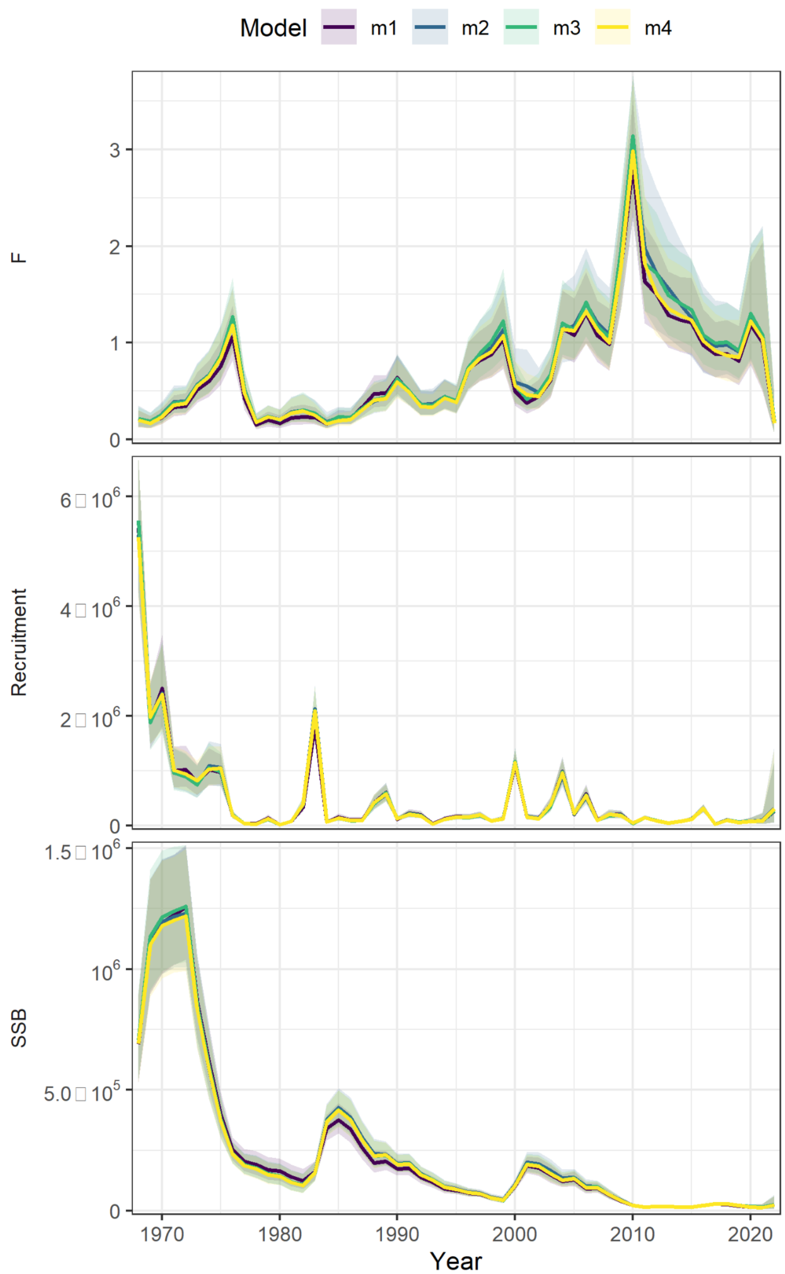

Figure 9: Comparison of runs with different time-varying selectivity assumptions (m1 = none; m2 = ar1_y; m3 = iid; m4 = Beverton holt).

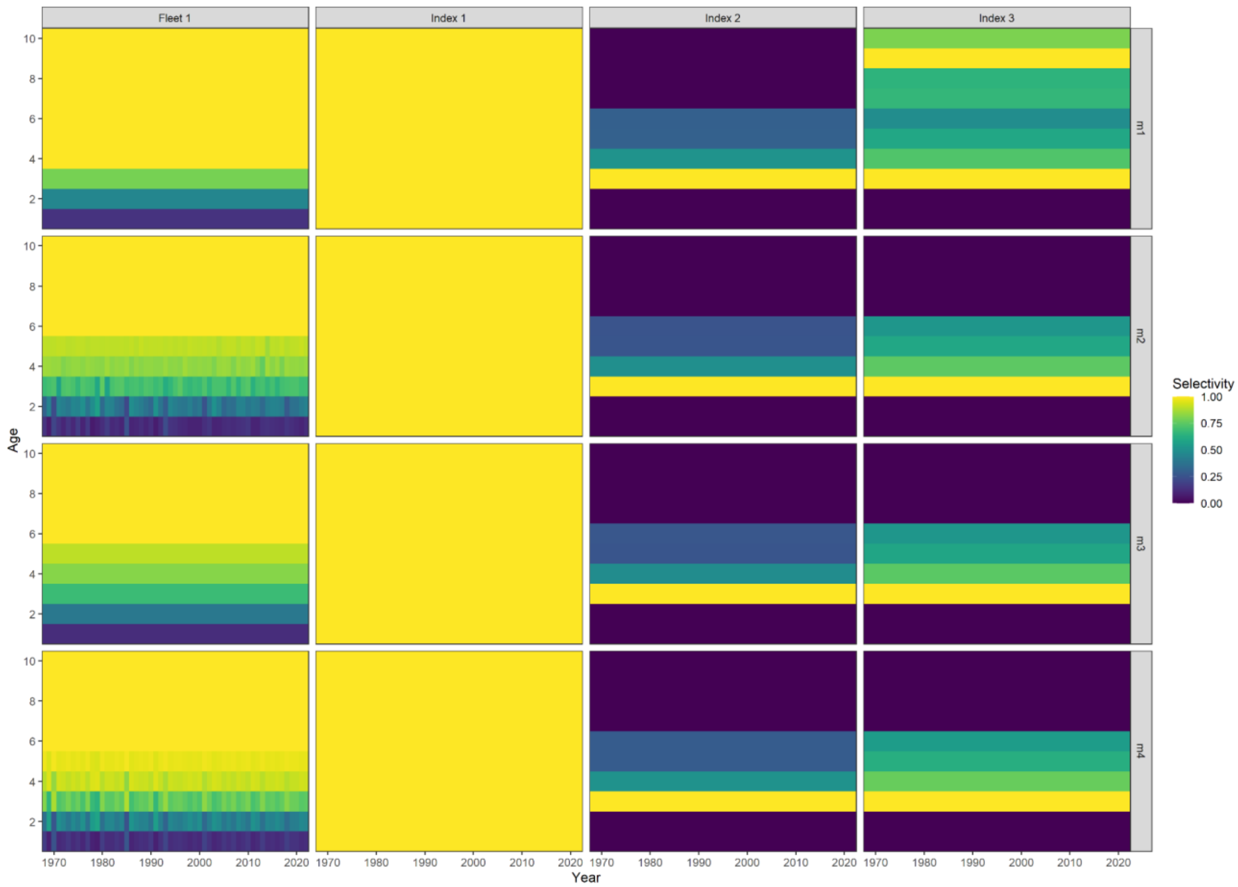

Figure 10: Comparison of time-varying selectivity in the fishing fleet across WHAM runs (m1 = none, m2 = iid, m3 = ar1_y, m4 = 2dar1).

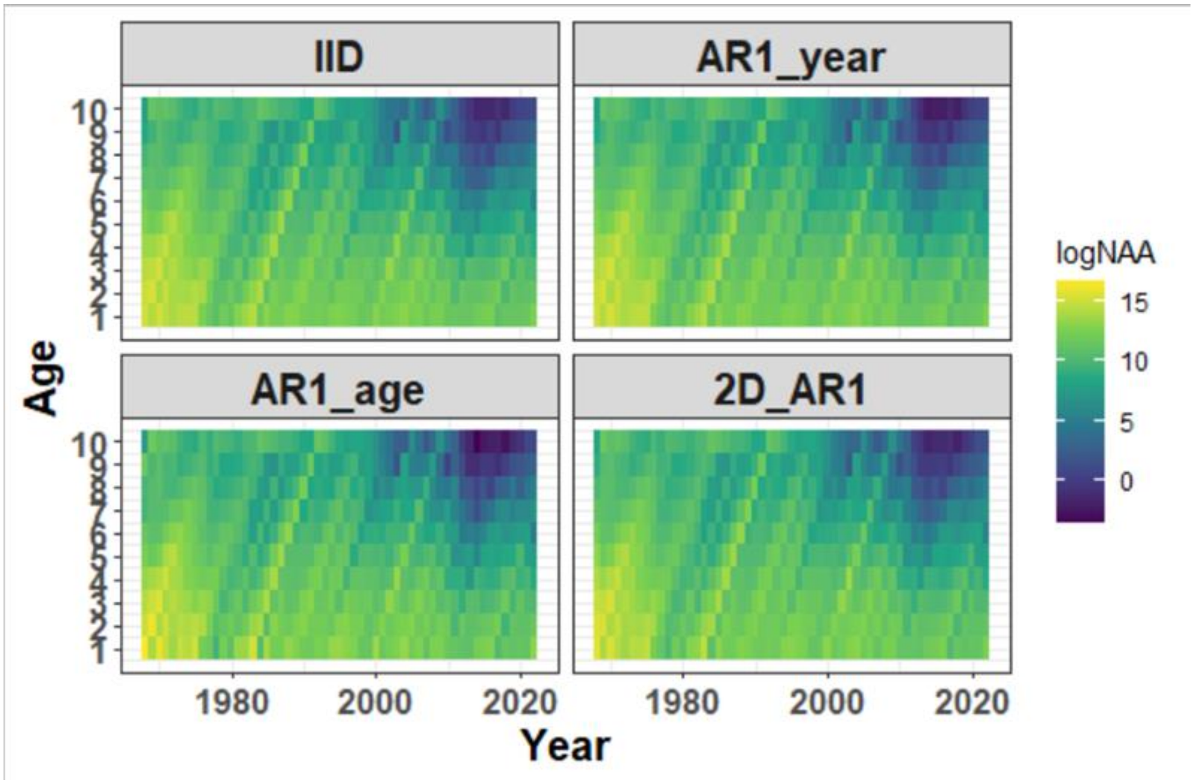

Figure 11: Comparison of estimated numbers-at-age among WHAM runs with varying random effects assumptions on numbers-at-age.

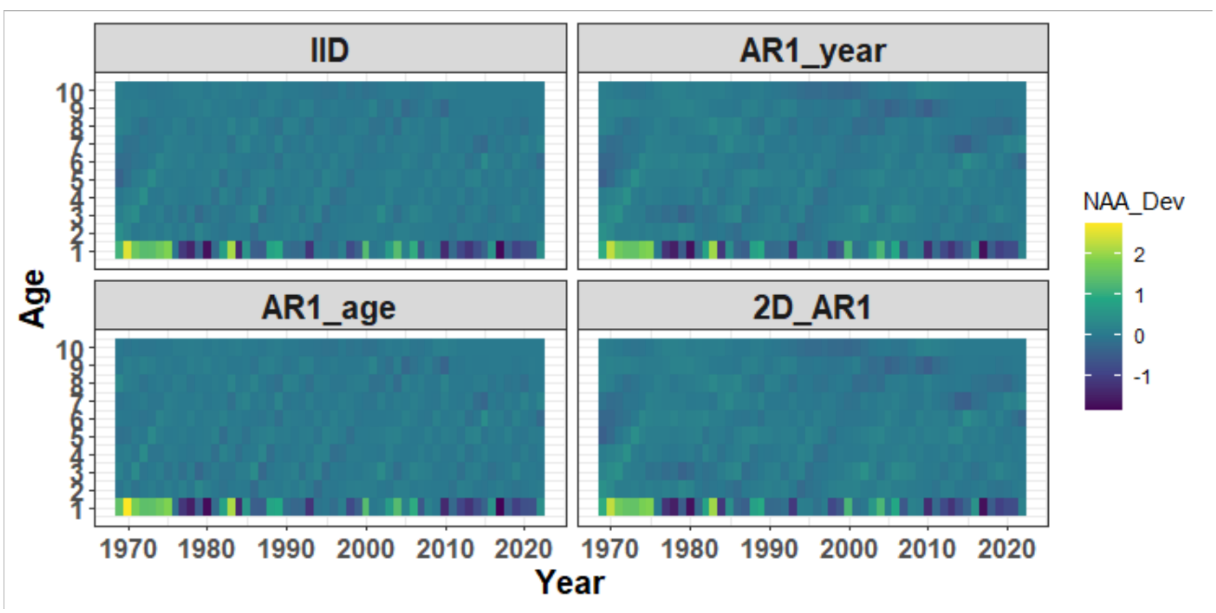

Figure 12. Comparison of deviations in numbers-at-age among WHAM runs with varying random effects assumptions on numbers-at-age.

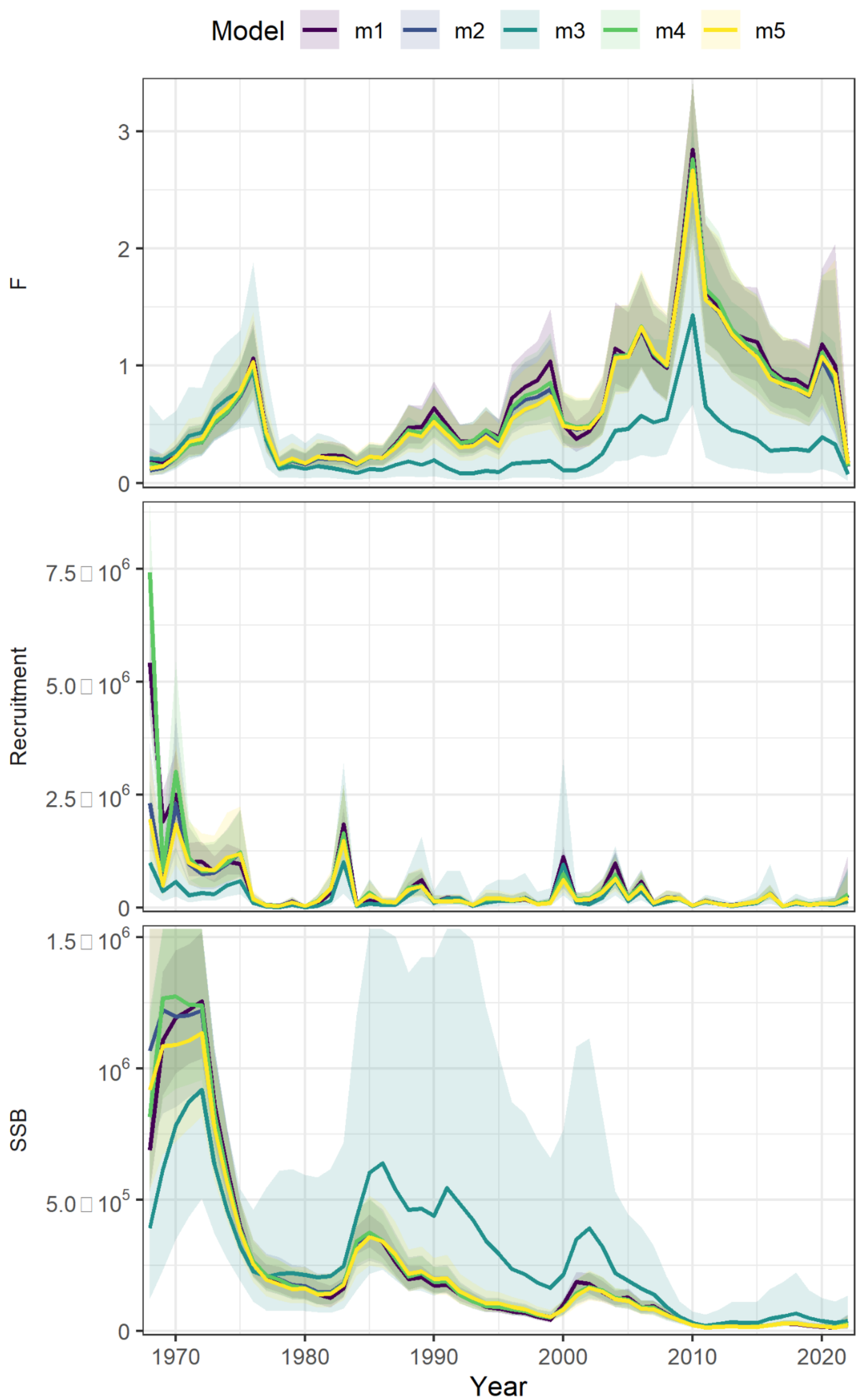

Figure 13: Comparison of runs with different random effects on numbers at age assumptions (m1 = none; m2 = iid; m3 = 2dar1; m4 = ar1_a, m5 = ar1_y).

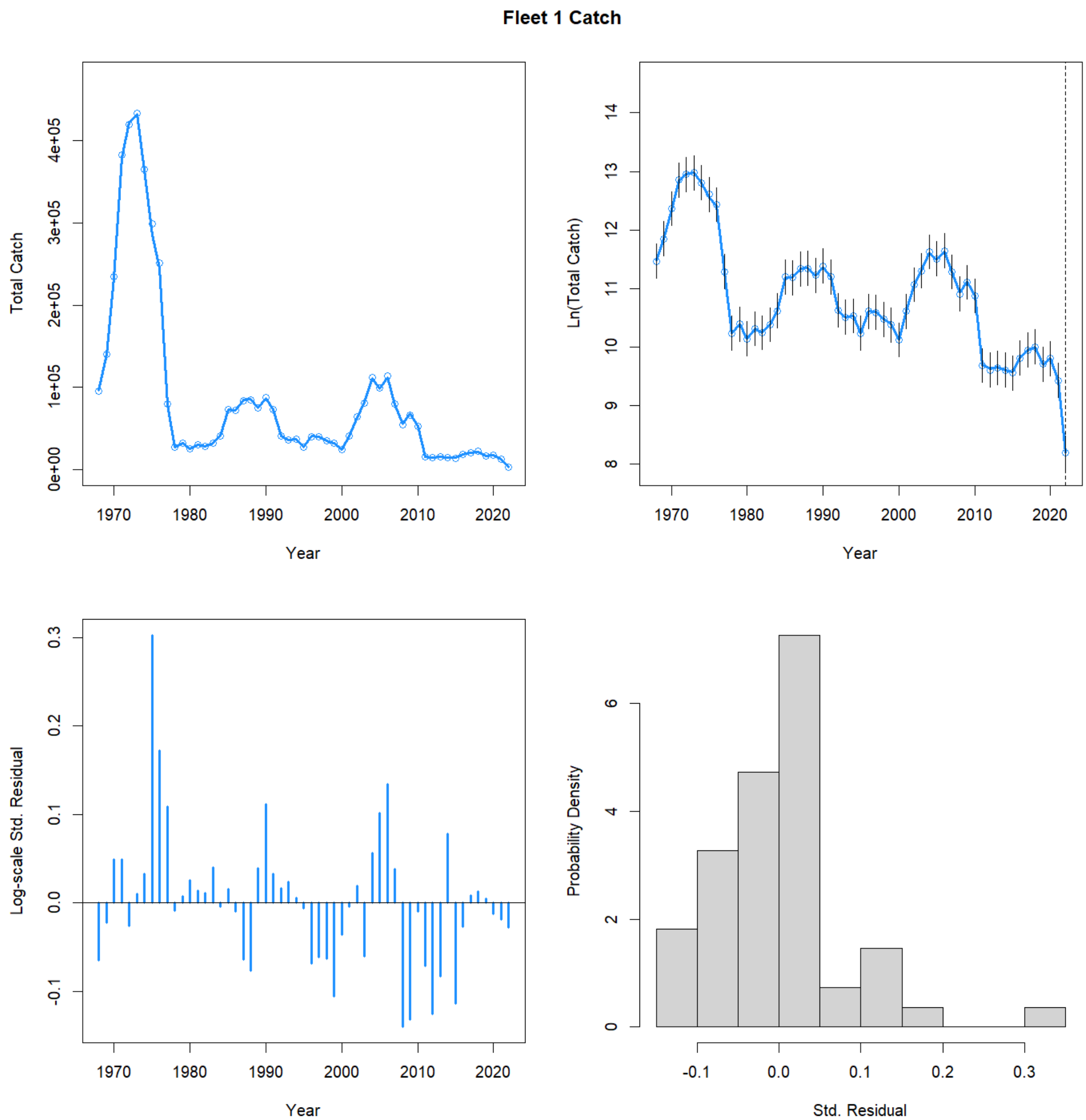

Figure 14: Fit of the proposed WHAM run to fishery catch.

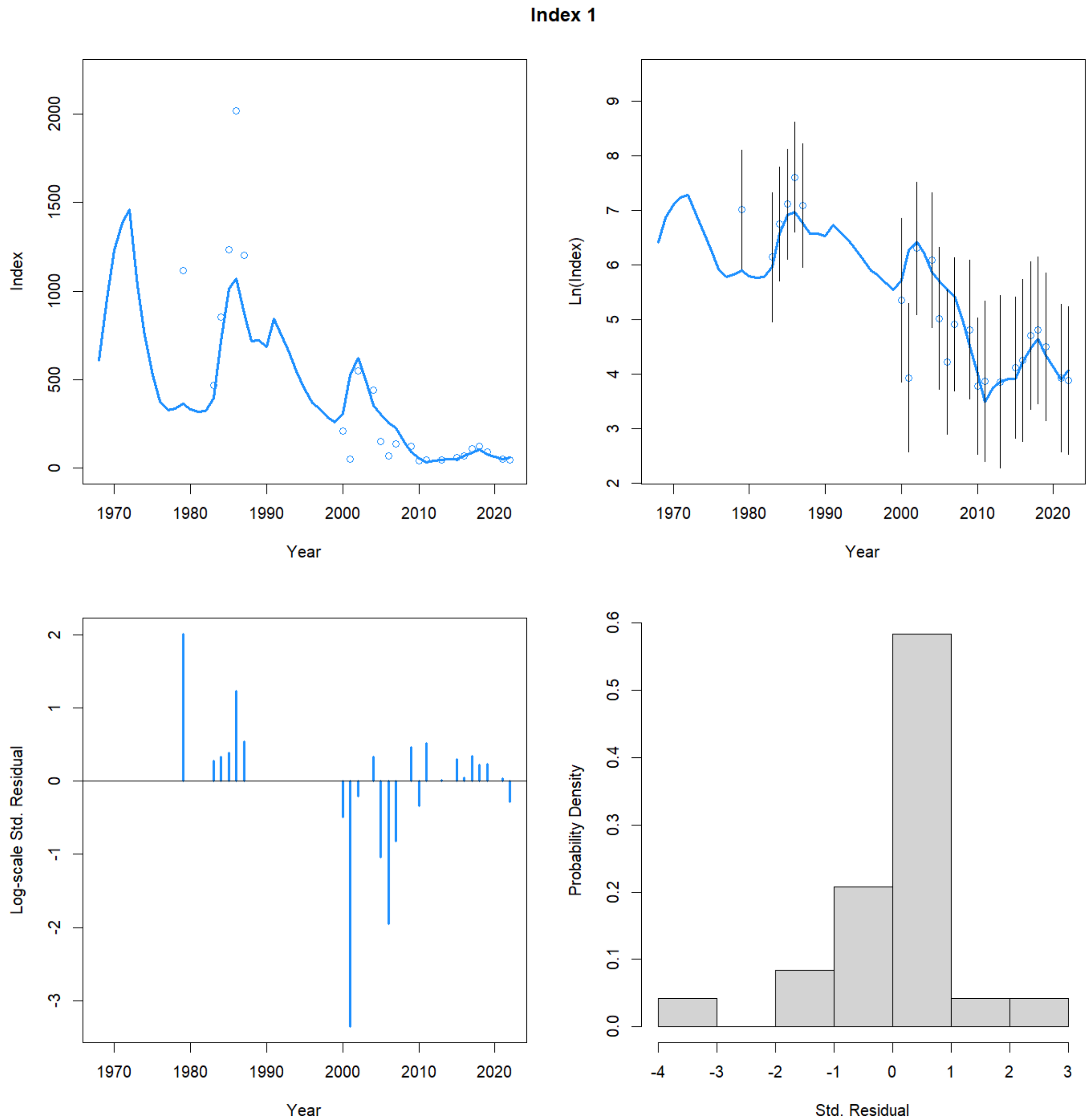

Figure 15: Fit of the proposed WHAM run to the rangewide egg index.

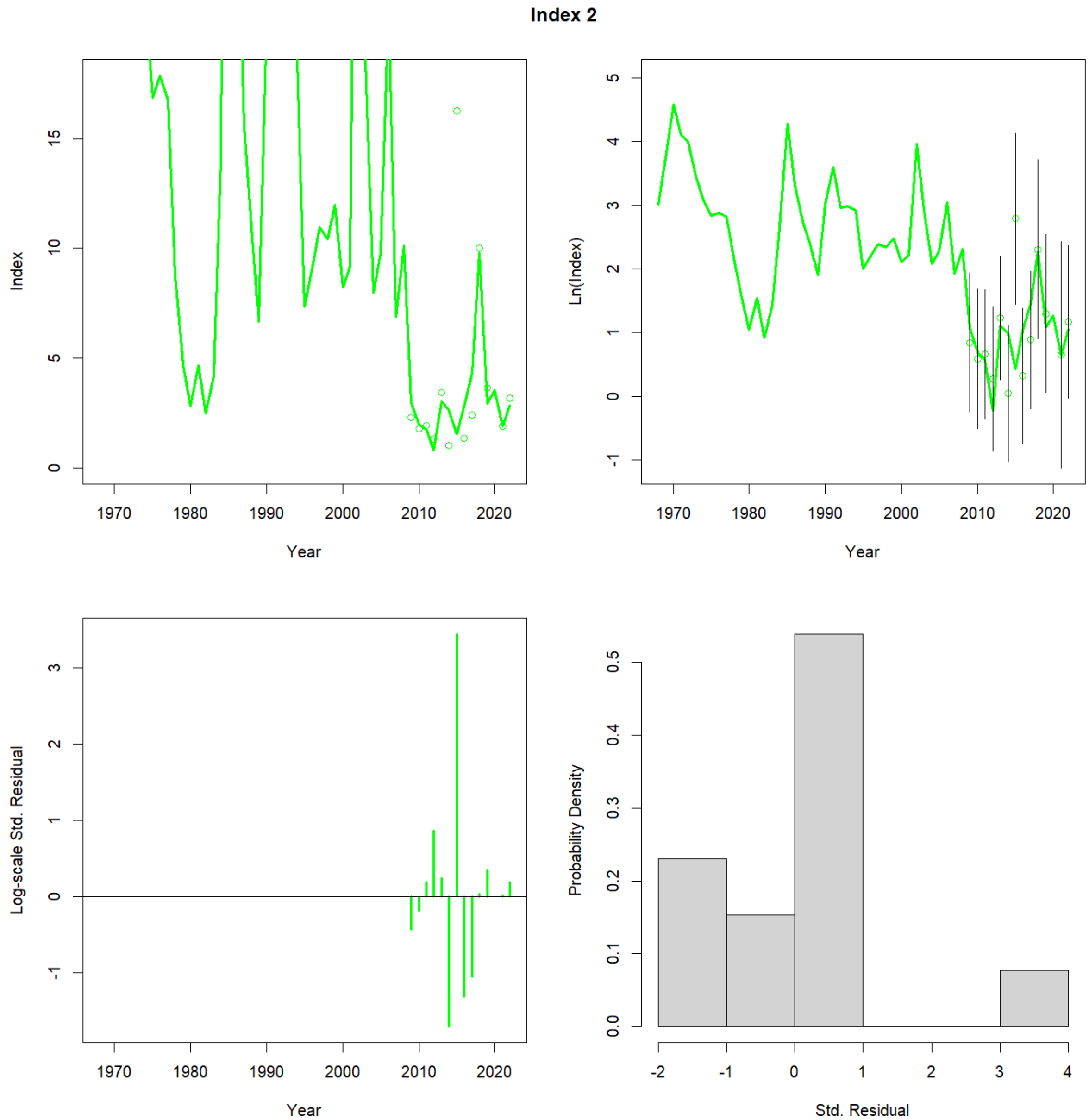

Figure 16: Fit of the proposed WHAM run to the NEFSC Bigelow survey.

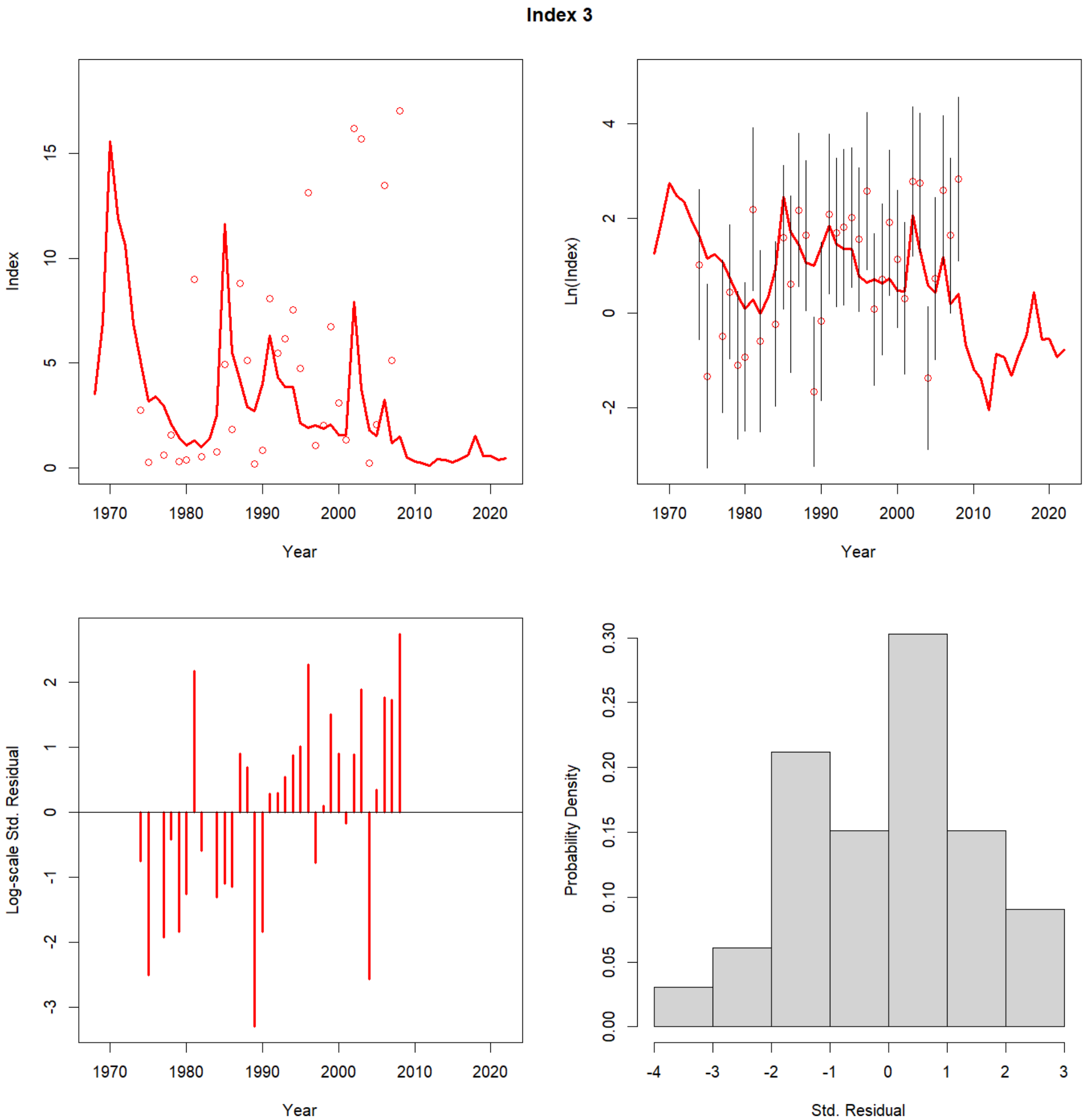

Figure 17: Fit of the proposed WHAM run to the NEFSC Albatross survey.

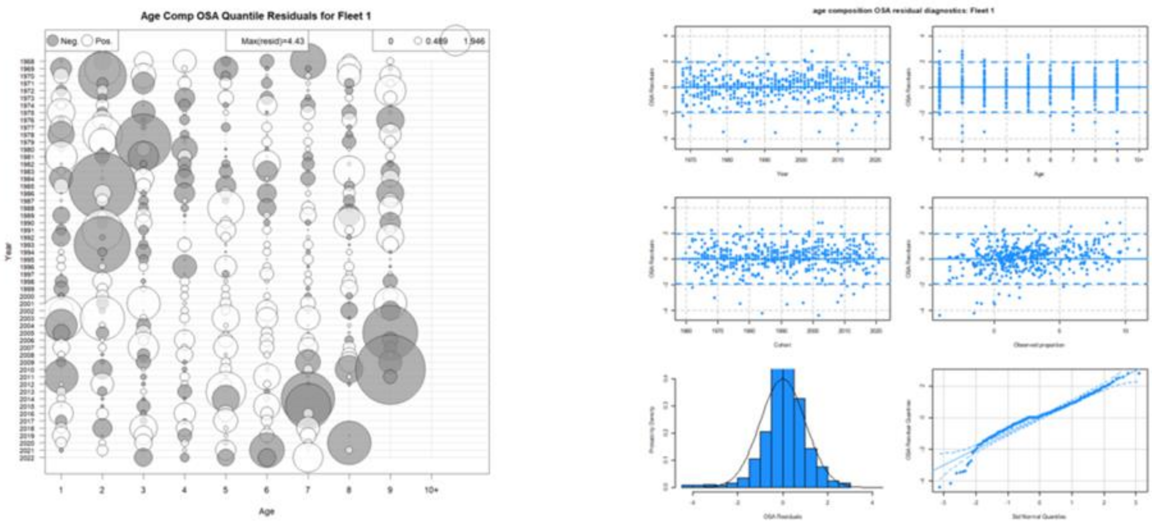

Figure 18: Logistic-normal-ar1-miss0 OSA residuals for the fishing fleet age composition data.

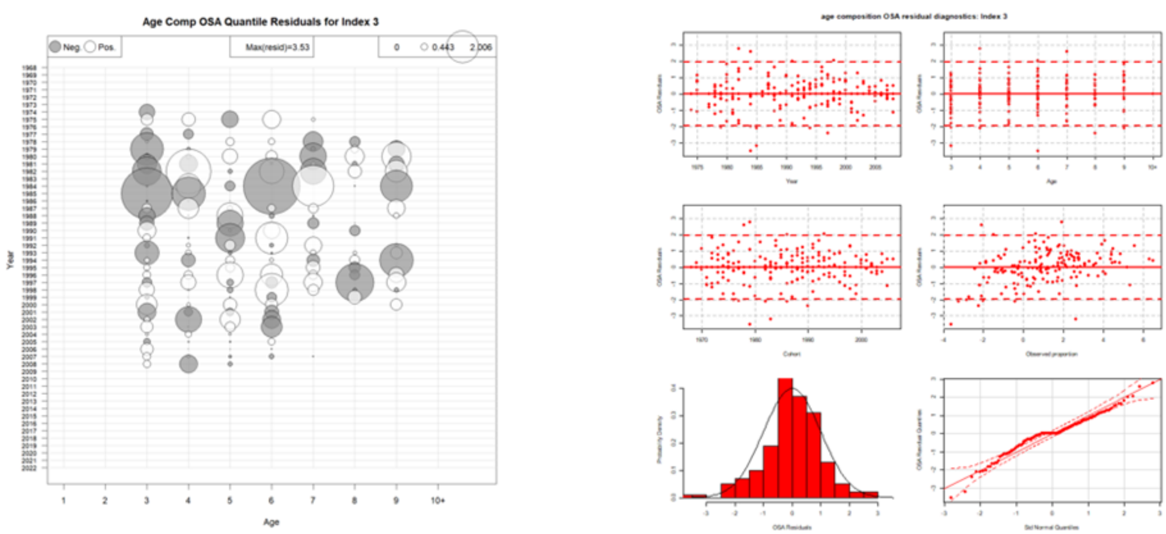

Figure 19: Logistic-normal-ar1-miss0 OSA residuals for NEFSC Albatross age composition data.

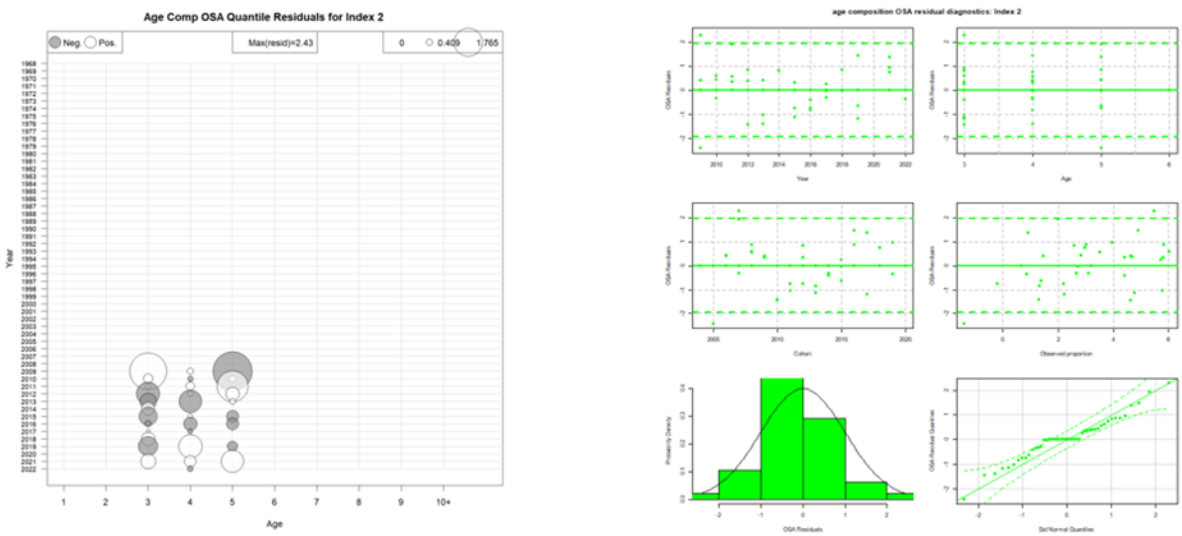

Figure 20: Logistic-normal-ar1-miss0 OSA residuals for NEFSC Bigelow age composition data.

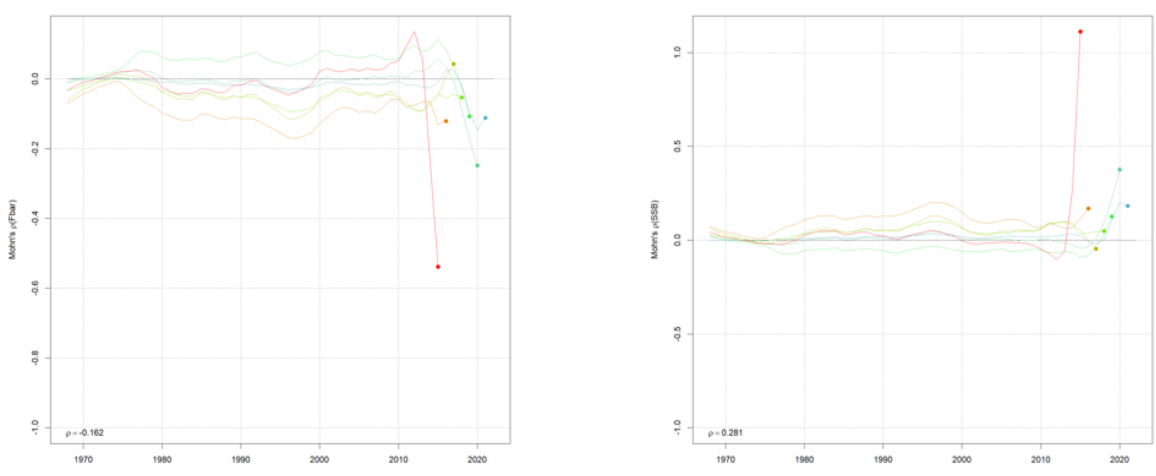

Figure 21: Retrospective patterns in fishing mortality and SSB for the proposed WHAM model.

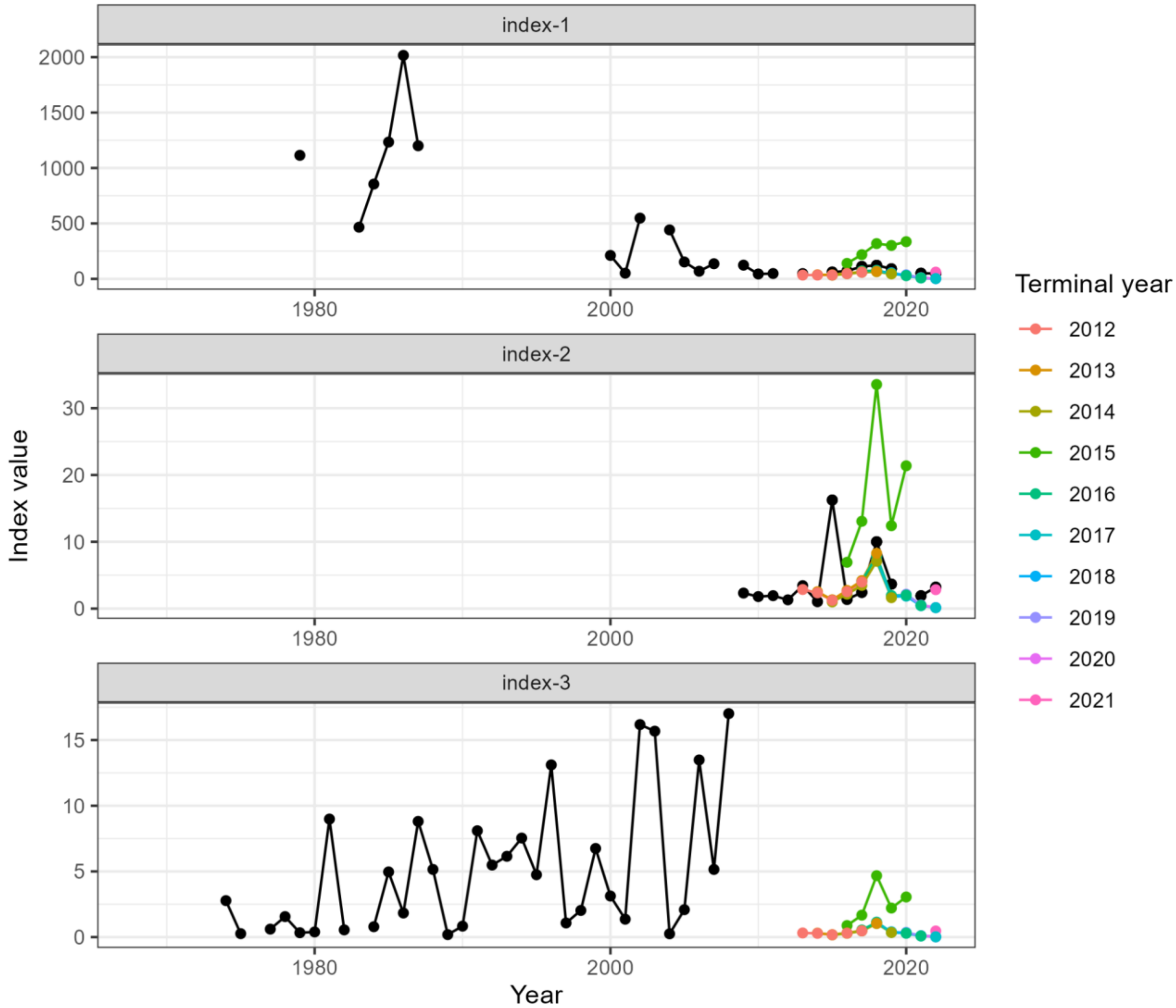

Figure 22: Predicted versus observed survey indices (index 1 = egg survey; index 2 = NEFSC Bigelow; index 3 = NEFSC Albatross). Colored points are predictions from the indicated terminal year.

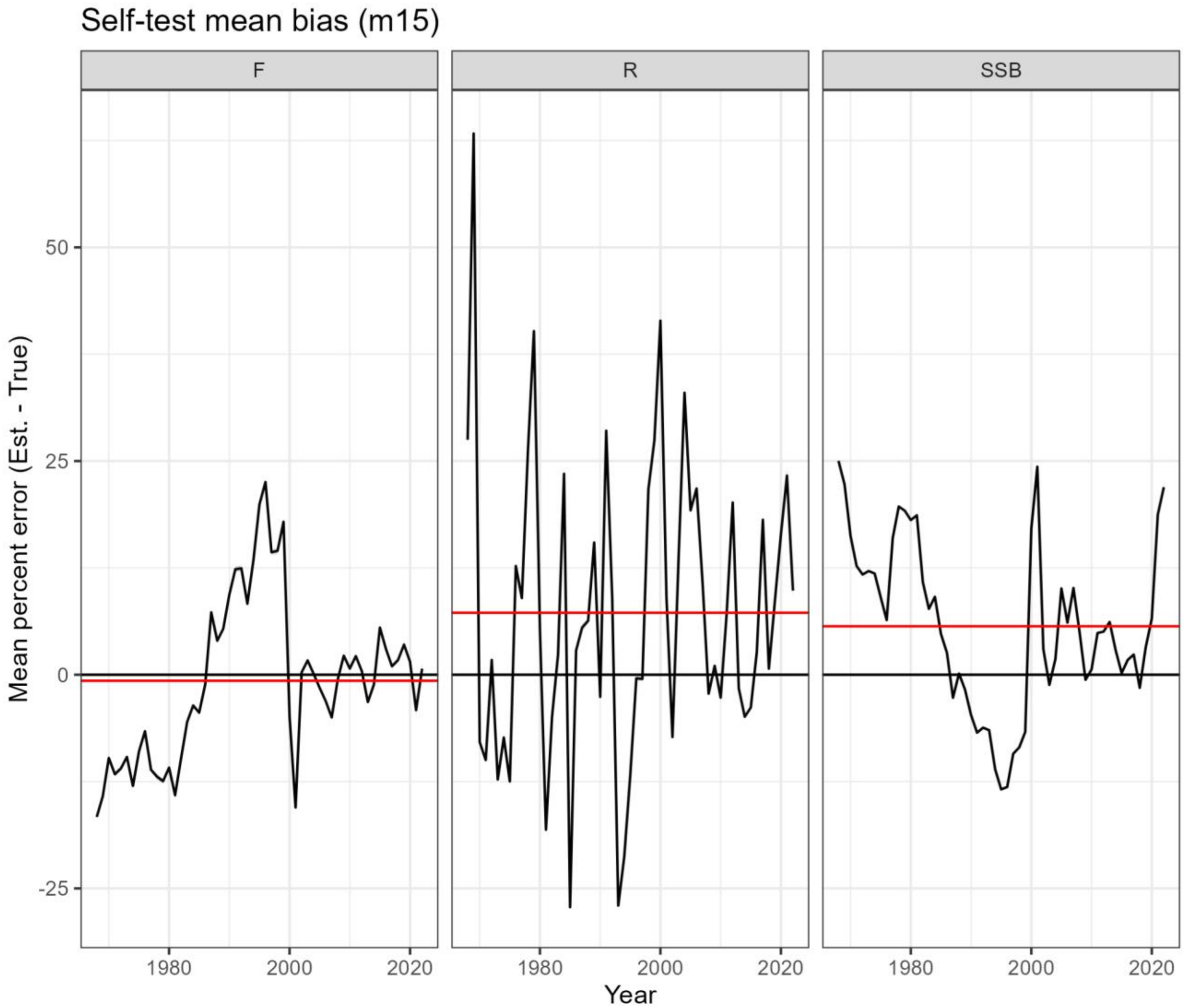

Figure 23: Time series of self-test mean percent bias for fishing mortality (F), recruitment (R) and spawning stock biomass (SSB).

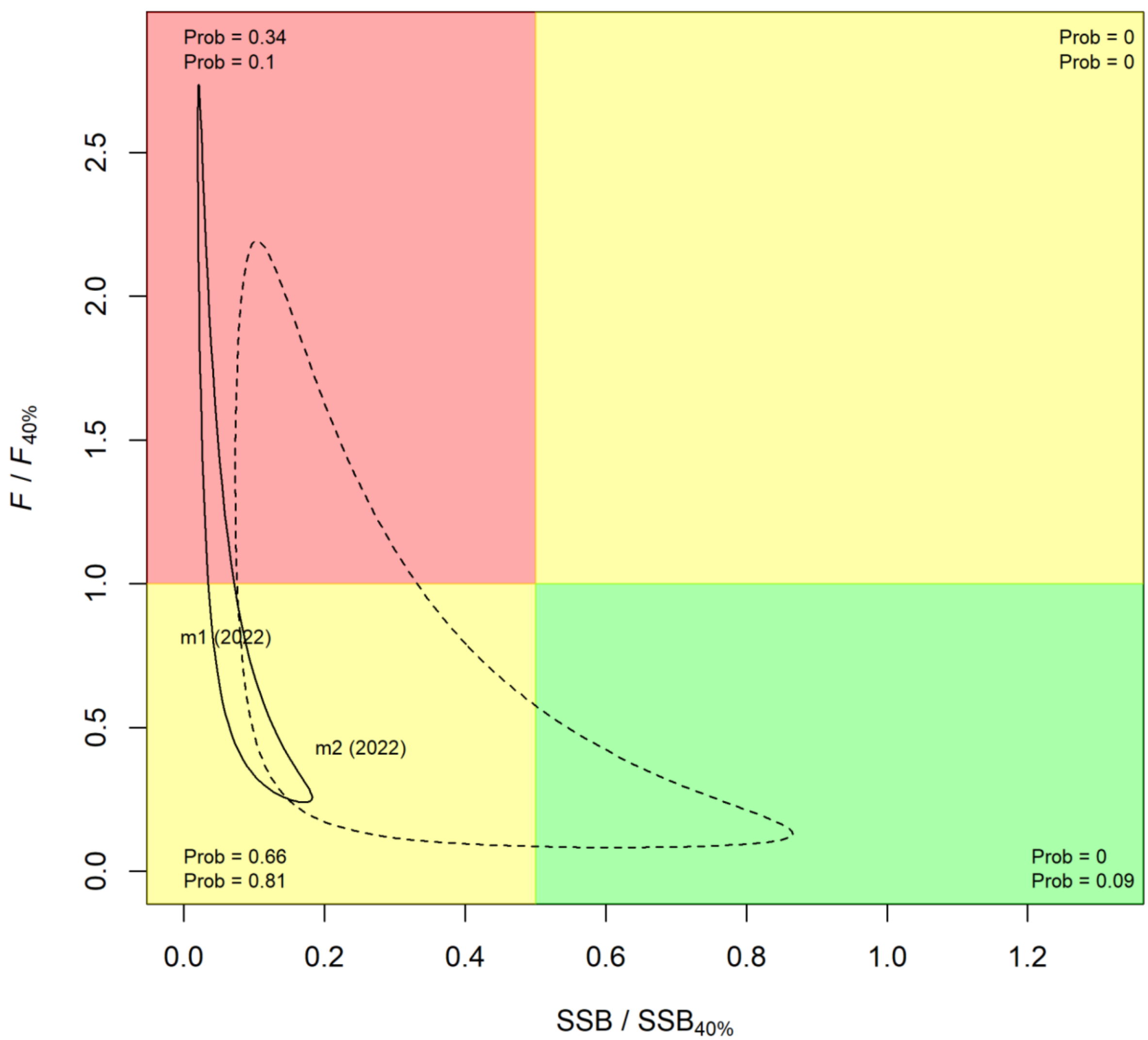

Figure 24: Comparison of stock status between the accepted ASAP (m1) and proposed WHAM (m2) models.

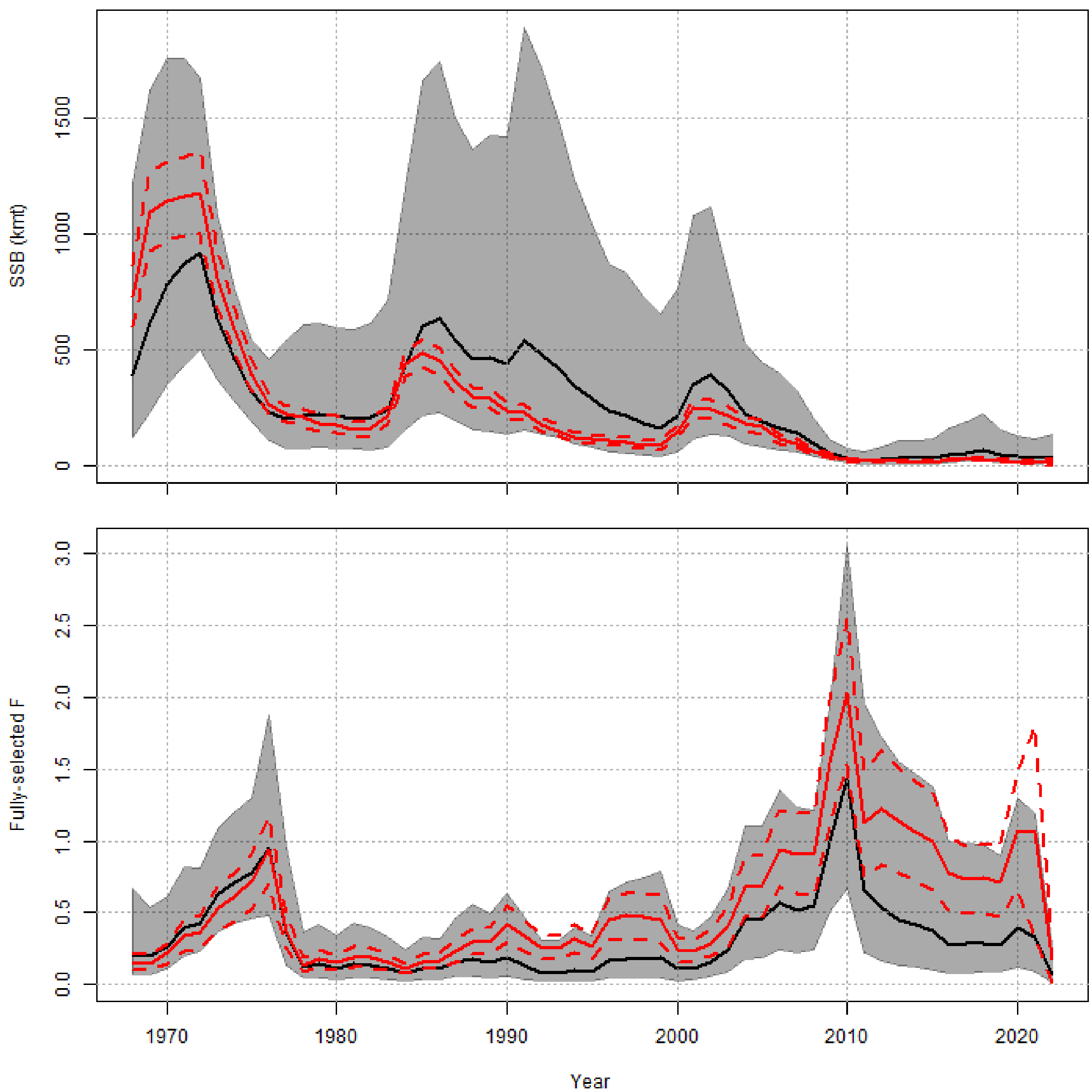

Figure 25: Comparison of trends in SSB and fishing mortality between the accepted ASAP (red) and proposed WHAM (black) models

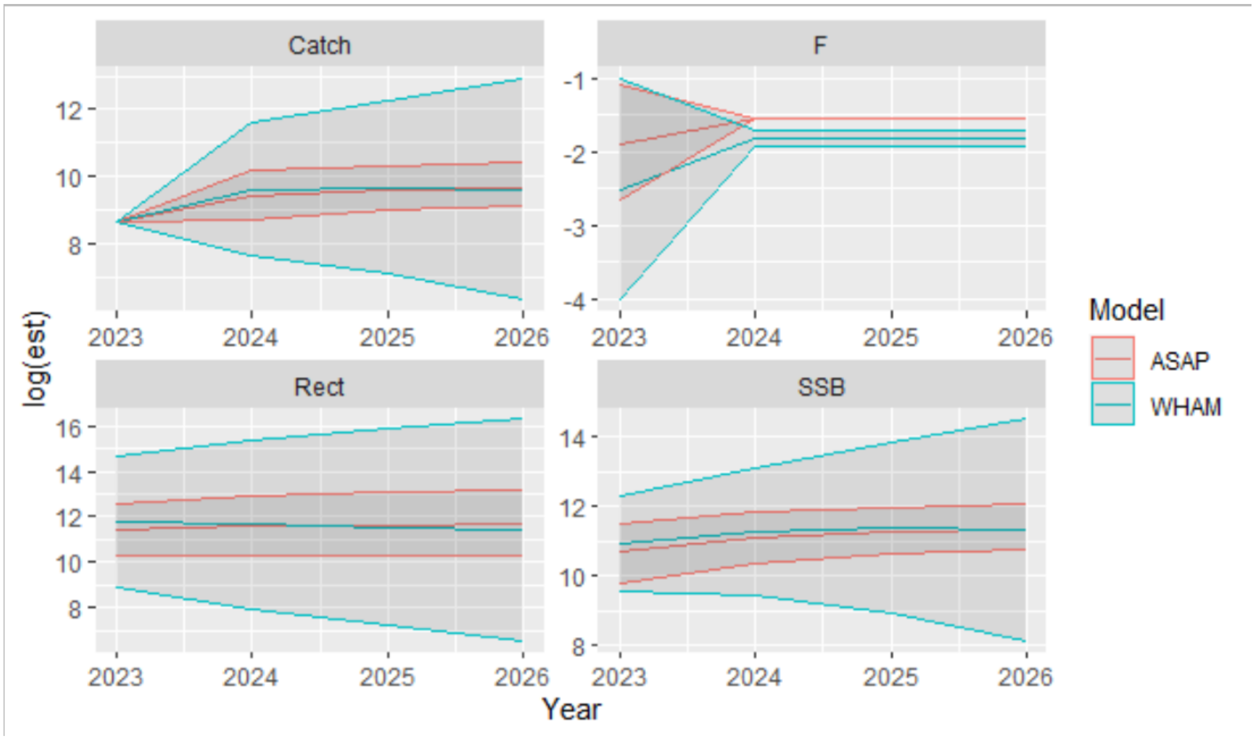

Figure 26: Comparison of ASAP and WHAM short-term projections. A bridge catch of 5,653 mt was used for 2023 and the F40% proxy was used for subsequent years.