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

Brian C. Stock[1], Timothy J. Miller[1]

November 2021

[1]brian.stock@noaa.gov, timothy.j.miller@noaa.gov, Northeast Fisheries Science Center, National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543, USA

Summary

The Woods Hole Assessment Model (WHAM) software package is being developed at the Northeast Fisheries Science Center to enable state-space stock assessments, i.e. where processes such as the annual transitions in numbers-at-age (NAA), *M*, and/or selectivity are treated as time- and age-varying random effects. WHAM can also be configured as a traditional statistical catch-at-age model in order to bridge from current assessments which use Age-Structured Assessment Program (ASAP).

We fit a series of models in WHAM for butterfish and consider three in detail. The simplest model, "04-Base," is similar to the final ASAP3 RUN_036 and estimates yearly recruitment as fixed effect parameters. The second model, "04-NAA2,"treats yearly recruitment deviations as random effects following a first-order autoregressive, AR(1), process. The proposed WHAM model, "17-NAA5," estimates all numbers-at-age (NAA) as random effects with 2D AR(1) covariance by age and year, but where only correlation by year is estimated. 17-NAA5 also uses the logistic normal likelihood for age composition observations. We compare diagnostics for these three models, and show that 17-NAA5 has higher prediction skill (of index observations) and higher convergence rate in simulation self-tests. We provide reference point calculations and short-term projections and propose that 17-NAA5 be used for butterfish management.

1 Introduction

Like most stocks with age-structured assessments in the U.S. Northeast, butterfish is currently assessed using ASAP, the Age-Structured Assessment Program (Legault and Restrepo, 1998; Miller and Legault, 2015). ASAP is a statistical catch-at-age (SCAA) model which typically only considers yearly fishing mortality (F_y) and recruitment (R_y) as time-varying parameters. Other parameters are assumed constant primarily because there are not usually enough degrees of freedom to estimate them as time-varying. ASAP can penalize the deviations, e.g. in recruitment as $R_y \sim \mathcal{N}(R_0, \sigma_R^2)$, although the penalty terms, σ_R^2, must be fixed or iteratively tuned and are therefore somewhat subjective (Aeberhard et al., 2018; Methot and Taylor, 2011; Methot and Wetzel, 2013; Xu et al., 2020). State-space models that treat parameters as unobserved states can, in principle, avoid such subjectivity by estimating the penalty terms as variance parameters constraining random effects and maximizing the marginal likelihood (Thorson, 2019). In this way, state-space models can allow processes to vary in time while simultaneously estimating fewer parameters. In addition to this key advantage, state-space models naturally predict unobserved states, and therefore handle missing data and short-term projections in a straightforward way (ICES, 2020). In comparisons with SCAA models, state-space models have been shown to have larger, more realistic, uncertainty and reduced retrospective patterns (Miller and Hyun, 2018; Stock et al., 2021; Stock and Miller, 2021).

The Woods Hole Assessment Model (WHAM) is an R package developed at the Northeast Fisheries Science Center (https://timjmiller.github.io/wham, Miller and Stock, 2020). It is similar to ASAP and can be configured to fit SCAA models nearly identically. There is functionality built into WHAM to migrate ASAP input files to R inputs needed for WHAM, and WHAM uses many of the same types of data inputs, such as empirical weight-at-age, so that existing assessments in the U.S. Northeast can be replicated and tested against models with state-space and environmental effects in a single framework. WHAM primarily differs from ASAP through inclusion of random effects options and implementation via the Template Model Builder package (TMB, Kristensen et al., 2016). In this respect it is similar to the State-space Assessment Model (SAM, Nielsen and Berg, 2014), which is currently used to manage roughly 25 stocks in the ICES region. To date, WHAM has not been used for management. However, it has been reviewed in the literature and simulation tested (Stock et al., 2021; Stock and Miller, 2021) and used as the operating model in a recently reviewed research track assessment (https://github.com/cmlegault/IBMWG, Legault et al., 2021). WHAM is also being considered in ongoing stock-specific research track assessments for Georges Bank haddock and American plaice.

Here, we describe a series of WHAM models for butterfish and consider three in detail. The simplest model, "04-Base," mimics ASAP and estimates yearly recruitment as fixed effect parameters. The second, "04-NAA2," treats yearly recruitment deviations as random effects following a first-order autoregressive, AR(1), process. The proposed WHAM model, "17-NAA5," estimates all numbers-at-age (NAA) as random effects with AR(1) correlation by year, but independent across ages. The 17-NAA5 model also assumes logistic normal likelihoods for catch and index age composition observations. We compare diagnostics for these three models, and show that 17-NAA5 has higher prediction skill (of index observations) and higher convergence rate in simulation self-tests. We provide reference point calculations and short-term projections and propose that 17-NAA5 be used for butterfish management.

2 Methods

Stock and Miller (2021) provide a complete description of the WHAM model equations, simulation tests for 5 stocks, and demonstrations of the random effects options. Source code, documentation, vignettes, automated tests, issue tracking, and development news are available at https://timjmiller.github.io/wham/.

2.1 Model configurations

We ran all WHAM models using the input data file from the final ASAP3 model, `RUN_036`. We investigated the following:

1. Numbers-at-age (NAA) model options
2. Estimating catchability (q) of Index 1 (NEFSC Fall Albatross)
3. Estimating natural mortality (M)
4. Age composition likelihood options
5. Estimating Beverton-Holt stock-recruitment
6. Time-varying selectivity vs. 2 blocks for the fishery

Table 5 lists all of the WHAM runs with description and comments. Code to run the final three WHAM models can be seen at `/code/run_models.R`. Code to extract reference point estimates and perform short-term projections is at `/code/project_models.R`.

2.1.1 Numbers-at-age models

Our notation for the NAA options follows Stock and Miller (2021). The "Base" model approximates ASAP by estimating recruitment deviations as independent fixed effect parameters. WHAM can also treat only recruitment (NAA1 and NAA2) or numbers at all ages (NAA3, NAA4, and NAA5) as random effects. Models with only recruitment as random effects are technically state-space models, and we therefore refer to models with all NAA as random effects as "full state-space" models, i.e. include process error on the NAA transitions (akin to "survival," Stock et al., 2021). Table 1 lists standard WHAM options for treating NAA as fixed or random effects.

Table 1: Six standard numbers-at-age (NAA) models in WHAM.

Model	Description	Parameters estimated	No.
Base	as ASAP, recruitment deviations are fixed effects	R_y for $y > 1$	$n_{years} - 1$
NAA1	Recruitment deviations are independent random effects	σ_R	1
NAA2	Recruitment deviations are autocorrelated, AR(1), random effects	σ_R, ρ_{year}	2
NAA3	All NAA deviations are independent random effects	σ_R, σ_a	2
NAA4	All NAA deviations are random effects with correlation by year and age, 2D AR(1)	σ_R, σ_a, ρ_{year}, ρ_{age}	4
NAA5	All NAA deviations are random effects with correlation by year only, AR(1)	σ_R, σ_a, ρ_{year}	3

We present results from Base, NAA2, and NAA5 models (the butterfish recruitment time-series exhibits strong autocorrelation by year). The full state-space models, NAA3–NAA5, did not converge with multinomial age composition likelihood but did with logistic-normal.

2.1.2 Estimating catchability (q) of Index 1 (NEFSC Fall Albatross)

Catchability of Index 1 (NEFSC Fall Albatross), q_1, is technically estimated in the ASAP3 model, `RUN_036`. However, the very strong penalty (CV = 0.01) results in the estimate, 0.197517, remaining close to the initial value, 0.21. We attempted to freely estimate q_1 in WHAM, i.e. without a penalty, but these models had issues estimating the population scale. Fixing q_1 at the value estimated in `RUN_036`, 0.197517, resulted in a lower negative log-likelihood than fixing q_1 at the `RUN_036` initial value, 0.21. Therefore, all three WHAM models presented fix $q_1 = 0.197517$.

2.1.3 Estimating natural mortality (M)

The ASAP3 model, `RUN_036`, fixes $M = 1.278$ for all ages. WHAM has several options for estimating M (see https://timjmiller.github.io/wham/articles/ex5_GSI_M.html), and we attempted to estimate a single mean M. Several of these models converged and generally estimated M lower than 1.278, in the 0.9-1.0 range with 95% CI from 0.6-1.4. These models estimated lower F, higher SSB, lower recruitment, and higher uncertainty in all three quantities. Selectivity was more domed for the indices and shifted younger for the fleet. Estimating M was not supported by AIC and had lower prediction skill, so we did not pursue these models further.

2.1.4 Age composition likelihood options

ASAP assumes that the age composition (proportion-at-age) observations follow the multinomial likelihood, where the effective sample size must be specified by the user. Although the multinomial is commonly used, it has two primary drawbacks: 1) the effective sample size weights the observations and cannot be estimated internally, and 2) the correlations are negative and completely defined by the mean of the distribution (Francis, 2014). WHAM provides several alternative composition likelihoods (Appendix B in Stock and Miller, 2021), including the Dirichlet-multinomial and logistic-normal, which have been shown to outperform the multinomial in simulation tests (Fisch et al., 2021; Francis, 2014; Thorson, 2019; Xu et al., 2020).

Models using the Dirichlet-multinomial did not converge, but models with the logistic-normal were promising. Of the three models presented in detail below, 04-Base and 04-NAA2 retain the multinomial from the ASAP3 model, whereas 17-NAA5 uses the logistic-normal.

2.1.5 Stock-recruitment

Estimating a stock-recruit function is desirable in part because it allows the use of MSY-based reference points. Ideally this would be done internally, within the model, but can also be done externally using estimated SSB and recruitment time-series. We were able to estimate Beverton-Holt parameters for some WHAM models. However, they are not appropriate because recruits in the butterfish assessment are age-0, and WHAM assumes age-1 recruits enter the population on Jan 1. Several modifications need to be made to allow for age-0 recruitment in WHAM, and there is no timeline for conducting this work. Thus, all three WHAM models presented assume recruitment deviations are random about the mean, R_0, with lognormal bias correction. This could be reevaluated in the future if 1) an age-0 recruitment option is developed in

WHAM, or 2) age-0 data are removed from the model (i.e. estimate age-1 recruits instead of age-0 recruits).

2.1.6 Time-varying selectivity vs. 2 blocks for the fishery

The final ASAP3 model, `RUN_036`, has a second selectivity block for the fishery from 2014-2019 (see results and justification for ASAP runs 32 and 33). An alternative to this 2-block structure in WHAM is to estimate time-varying selectivity deviations as random effects (see https://timjmiller.github.io/wham/articles/ex4_selectivity.html). We fit WHAM models with time-varying age-specific and logistic selectivity parameters. Models with time-varying logistic selectivity did not converge, which is unsurprising given the reasonably strong doming when age-specific selectivity is estimated. One model with time-varying age-specific selectivity was promising but did not have better diagnostic performance than the proposed WHAM model, 17-NAA5. Models with random effects on both selectivity and all NAA did not converge.

2.2 Diagnostic and performance metrics

We primarily considered the following diagnostic and performance metrics:

- Convergence
- Trend in Index 1 residuals
- Akaike information criterion (AIC)
- Retrospective pattern (Mohn's ρ)
- Simulation self-test
- Predictive skill

2.2.1 Convergence

We considered models converged if 1) the minimization algorithm, `stats::nlminb`, indicated successful completion (`convergence = 0`), and 2) the Hessian was positive definite and standard errors were calculated for all parameters.

2.2.2 Trend in Index 1 residuals

Some models did not fit Index 1 (NEFSC Fall Bottom Trawl Survey (BTS), Albatross years 1989-2008) well, which can be seen as a trend in the Index 1 residuals. Fits to Indices 2-6 were adequate for all models and therefore not helpful in model selection. State-space models should be diagnosed using one-step ahead (OSA) residuals, which are conditioned on previous data points and independent (Berg and Nielsen, 2016; Thygesen et al., 2017).

2.2.3 Akaike information criterion (AIC)

WHAM calculates the marginal AIC, which is a useful model selection metric in some cases. Unfortunately, it cannot be used to select between models with different likelihood functions, e.g. multinomial versus logistic-normal age compositions, 17-NAA5 vs. 04-Base or 04-NAA2. It also cannot be used to compare models that treat the same parameters as fixed versus random effects, e.g. 04-Base vs. 04-NAA2. Therefore, while AIC was useful in some instances, it is not applicable to compare the three WHAM models presented in detail here.

2.2.4 Retrospective pattern (Mohn's ρ)

We used the WHAM default of 7 peels to calculate Mohn's ρ for recruitment, SSB, and fully-selected F. In addition to the Mohn's ρ values, we considered the pattern of the peels. Absolute values of Mohn's ρ less than 0.2 are not generally considered problematic. Confidence intervals to statistically test whether Mohn's $|\rho|$ are greater than 0 or different between models would be ideal but this is an open research question for state-space models.

2.2.5 Simulation self-test

We ran simulation self-tests by using each model to simulate 100 datasets keeping all fixed effect parameters at the MLEs. We then refit the models to these simulated datasets and calculated the convergence rate and relative error in SSB, F, recruitment, and predicted catch.

2.2.6 Predictive skill

Performance in hindcasts, or "model-free validation," can be used more generally than AIC, e.g. regardless of the likelihood or treatment of parameters as fixed or random effects. Predictive skill is also a desirable metric because it focuses on the accuracy of future, instead of historical, estimates of stock status and is therefore more relevant to management. In addition, removing and predicting data is arguably more informative than relying on diagnostics such as residual patterns, which "can be removed by adding more parameters than justified by the data," or retrospective patterns, which can be "removed by ignoring the data" (Carvalho et al., 2021; Kell et al., 2021).

We ran hindcasts by sequentially removing aggregate and age composition observations for one index at a time, re-fitting the models, and predicting the removed data. We calculated the mean absolute scaled error (MASE) of the predictions over time horizons used to provide butterfish management advice, e.g. 1-3 years. MASE < 1 means that the model is better than the naive/baseline forecast, and MASE $= 0.5$ means that model forecasts are 2x as accurate as naive/baseline.

2.3 Reference points and status determination

We calculated $F_{50\%SPR}$ and $B_{50\%SPR}$ internally in WHAM according to the working group's proposed assumptions: 1) average recruitment since 2011 (2011-2019), and 2) average SSB per recruit inputs (i.e. selectivity-, maturity-, and weight-at-age) over the last five model years (2015-2019).

WHAM can propagate uncertainty in model parameters into uncertainty in $F_{X\%SPR}$ and $B_{X\%SPR}$, and then into stock status. WHAM also includes estimates of covariance of $F/F_{X\%SPR}$ and $B/B_{X\%SPR}$. Here, we have extracted the MLEs for $F_{50\%SPR}$ and $B_{50\%SPR}$, but without estimates of uncertainty, as is current practice. Thus, we do not provide 95% CI for $F_{50\%SPR}$ and $B_{50\%SPR}$, and the uncertainty in $F_{2019}/F_{50\%SPR}$ and $B_{2019}/B_{50\%SPR}$ results from uncertainty in F_{2019} and B_{2019} alone. We can include uncertainty estimates for $F_{50\%SPR}$ and $B_{50\%SPR}$ if desired.

2.4 Projections

WHAM has several options for handling short-term projections internally. Code to run short-term projections for the final three WHAM models can be seen at `/code/project_models.R`.

Projections under alternative F for catch advice will be done in the upcoming management track assessment using data through 2021. Here we simply demonstrate how this would be done using WHAM. We show three alternative F scenarios over a 3-year projection period: $F = 0$, $F = F_{2019}$ (terminal year F / status quo), and $F = F_{50\%}$ (F_{MSY} proxy).

In the models that assume the NAA deviations follow an AR(1) process (04-NAA2 and 17-NAA5) we continued the process into the projection period for consistency. We note that assumptions in the short-term projections are distinct from those defining reference points, which should reflect expected stock productivity in the long-term. Continuing the AR(1) process is an objective way of projecting numbers at age that are correlated with those in the terminal year but that correlation dampens with increased projection years. The rate of dampening depends on the correlation parameter, ρ_{year}.

04-Base treats recruitment in the model years as fixed effect parameters, as in ASAP. WHAM then treats recruitment in the projection years, $\log(R_y)$, as random effects following:

$$\log(R_y) \sim \mathcal{N}(\mu, \sigma^2)$$

where μ and σ are the mean and standard deviation of $\log(R_y)$ calculated from a specified subset of model years. Here, we calculate μ and σ from 2011-2019 recruitment as in the reference point definition.

3 Results

3.1 Convergence

04-Base, 04-NAA2, and 17-NAA5 each converged with positive definite Hessian and maximum gradient < `1e-11`.

3.2 Index 1 residuals

04-Base did not exhibit a trend in the Index 1 residuals (NEFSC Fall BTS, Albatross years 1989-2008). 04-NAA2 and 17-NAA5 had mild, insignificant trends. Some models, e.g. 25-NAA4-FAA, did not fit Index 1 well, resulting in a significant residual trend and we removed them from consideration (Fig. 1).

3.3 Retrospective pattern

Mohn's $|\rho|$ values for F, recruitment, and SSB were 0.11 or less for all three models (Fig. 2). The last peel (to 2013) is worse than others, likely because the second fleet selectivity block begins in 2014. For all diagnostics plots, see `/results/model-name`.

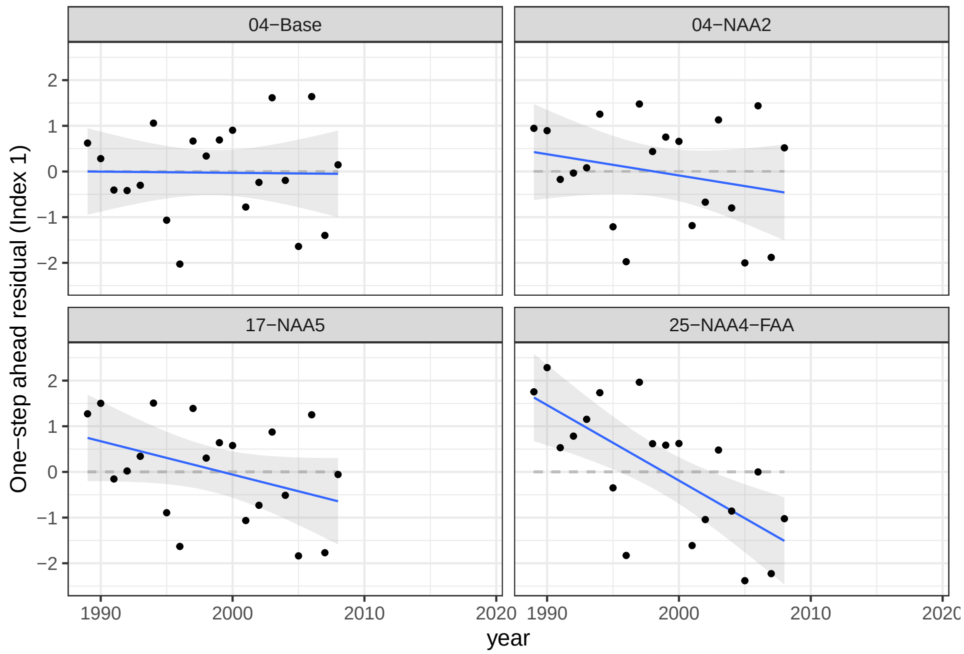

Figure 1: Trends in Index 1 one-step ahead (OSA) residuals.

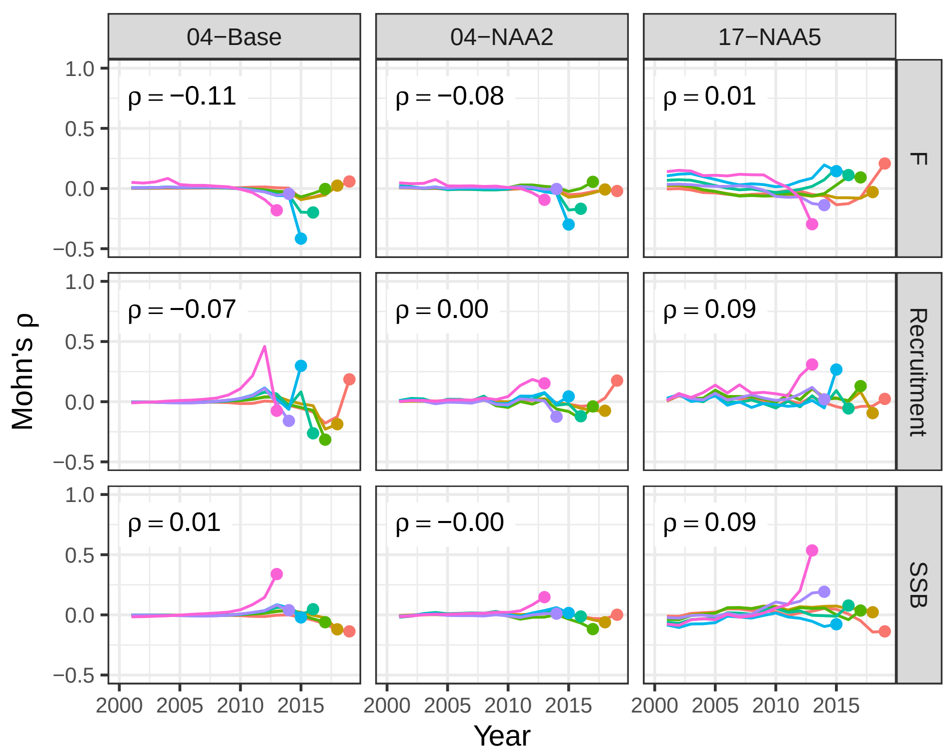

Figure 2: Retrospective patterns in F, recruitment, and SSB.

3.4 Numbers-at-age

The three final models primarily differ in their assumptions about the NAA transitions (Table 2, Fig. 3). 04-Base estimates annual recruitment deviations as independent fixed effect parameters. 04-NAA2 assumes recruitment is an AR(1) process, which smooths and reduces the magnitude of the deviations. 17-NAA5 is a full state-space model that allows for deviations in the NAA transitions at all ages with covariance by year. 04-NAA2 and 17-NAA5 estimated positive autocorrelation by year ($\rho_{year} > 0$, Table 2), which means that the negative recruitment deviations estimated in the terminal year propagate into the short-term projections (Fig. 3). 17-NAA5 estimated slightly positive survival deviations at ages 1+ in recent years, and these also propagate into the projections. Cohort effects can be seen in the NAA deviations estimated by 17-NAA5 (diagonal correlation in Fig. 3). At present, WHAM does not allow for cohort effects on the NAA deviations, but these could be considered in the future if added to WHAM.

Table 2: Maximum likelihood estimates of numbers-at-age (NAA) parameters in the three final WHAM butterfish models. Standard errors are in parentheses.

Model	σ_R	σ_a	ρ_{year}
04-Base	—	—	—
04-NAA2	0.17 (0.05)	—	0.88 (0.19)
17-NAA5	0.32 (0.07)	0.22 (0.06)	0.43 (0.21)

3.5 Selectivity

Fleet selectivity was estimated similarly in the three WHAM models as in ASAP `RUN_036` (Fig. 4).

Index selectivity was also estimated similarly in the three WHAM models, except that 17-NAA5 estimated lower selectivity of older butterfish, i.e. more doming (Fig. 5). Selectivity-at-age parameters for older ages (age-3 and especially the plus-group, age-4+) were not well estimated by 04-Base and 04-NAA2 (SE of several logit-scale parameters > 3), whereas they were for 17-NAA5 (maximum SE = 1.5).

3.6 Simulation self-test

When fit to data simulated from the fit models and keeping fixed effect parameters at their MLEs, 04-Base and 04-NAA2 converged less than half of the time. Convergence rates for 04-Base, 04-NAA2, and 17-NAA5

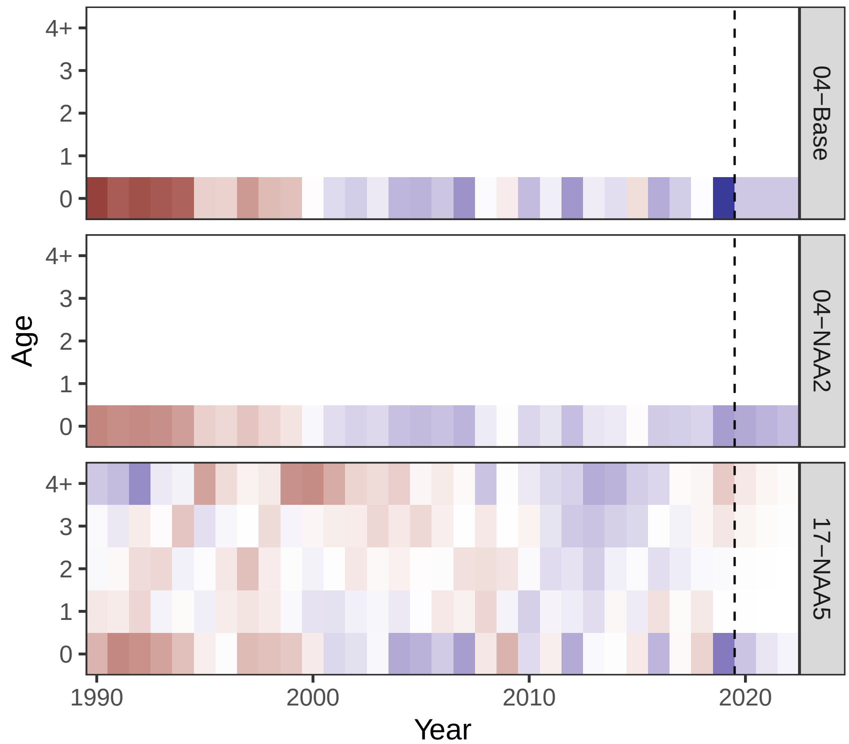

Figure 3: Numbers-at-age (NAA) deviations in the three final WHAM models. Positive and negative deviations are red and blue, respectively. Vertical dashed line indicates the terminal assessment year, 2019.

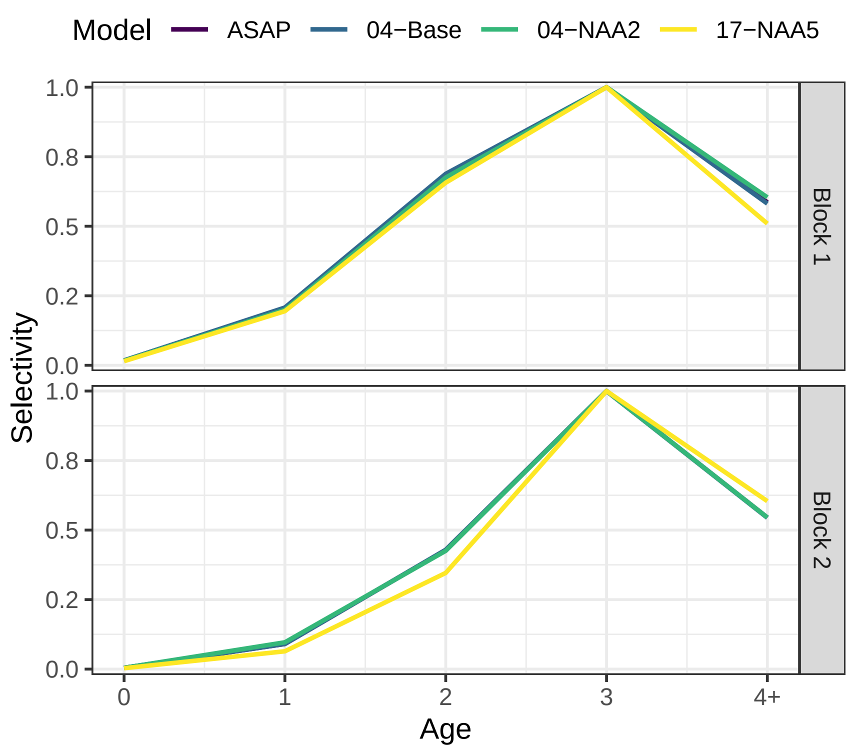

Figure 4: Fishery selectivity from ASAP RUN-36 and the three final WHAM models. Block 1: 1989-2013. Block 2: 2014-2019.

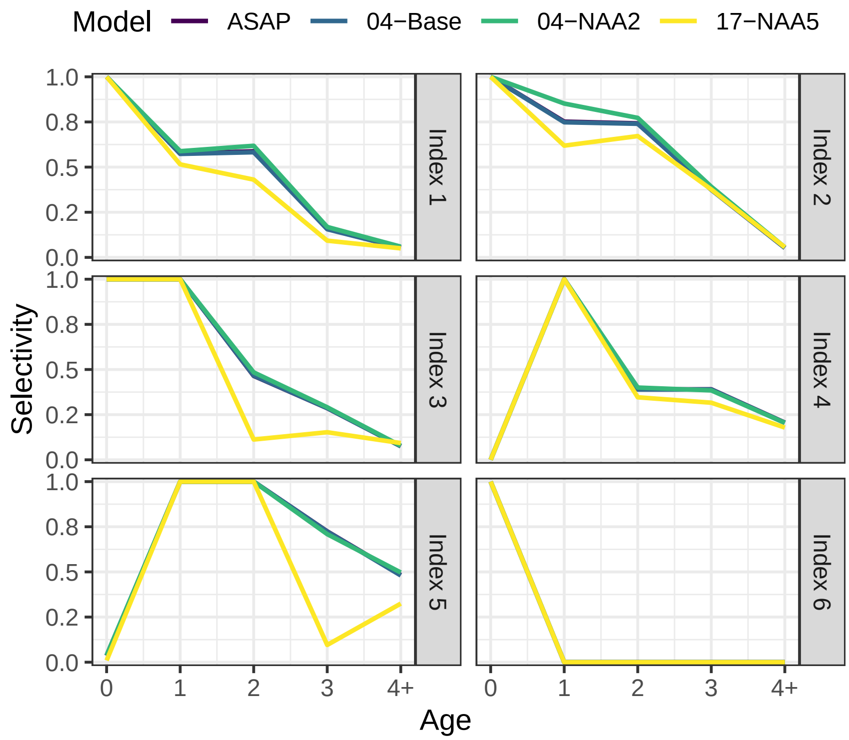

Figure 5: Index selectivity from ASAP RUN-36 and the three final WHAM models.

were 8%, 40%, and 95% respectively. None of the models exhibited bias in SSB, F, recruitment, or predicted catch (Fig. 6).

3.7 Predictive skill

Over time horizons used to provide butterfish management advice, i.e. 1-3 years, 04-NAA2 and 17-NAA5 had slightly higher predictive skill than 04-Base (lower median MASE, Fig. 7). All three models generally had MASE < 1, which means that they provide more accurate forecasts than the baseline (assumes index observation in following year will be the same as previous). The exceptions were Index 4 (NEFSC Spring, Bigelow years 2009-2019) at 2-year horizon and Index 6 (young of the year survey from combined state data) at 3-year horizon. Across all models and time horizons, prediction skill was highest for Indices 2 (NEFSC Fall, Bigelow years 2009-2019), 3 (NEAMAP Fall), and 5 (NEAMAP Spring), lowest for Index 6, and variable for Index 4.

3.8 Model selection

We recommend 17-NAA5 because it had a higher convergence rate in simulation self-tests and slightly higher median predictive skill (Table 3). We did not investigate all simulation fits, but we hypothesize that 04-Base

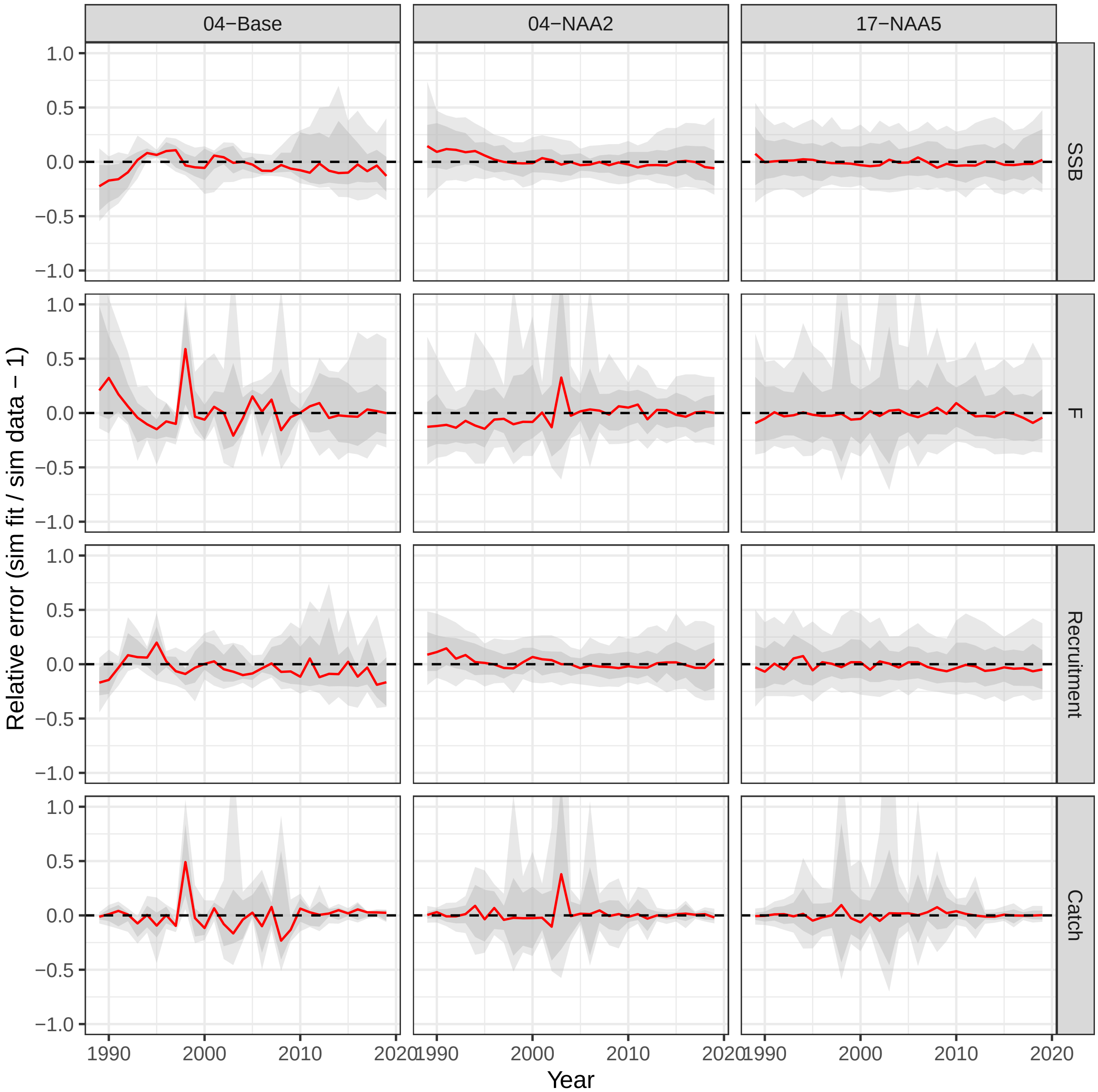

Figure 6: Simulation self-test results for the three WHAM models. Light grey shading shows the middle 80%, dark grey shows the middle 50%, and red lines are the medians of 100 simulations. Convergence rates for 04-Base, 04-NAA2, and 17-NAA5 were 8%, 40%, and 95% respectively.

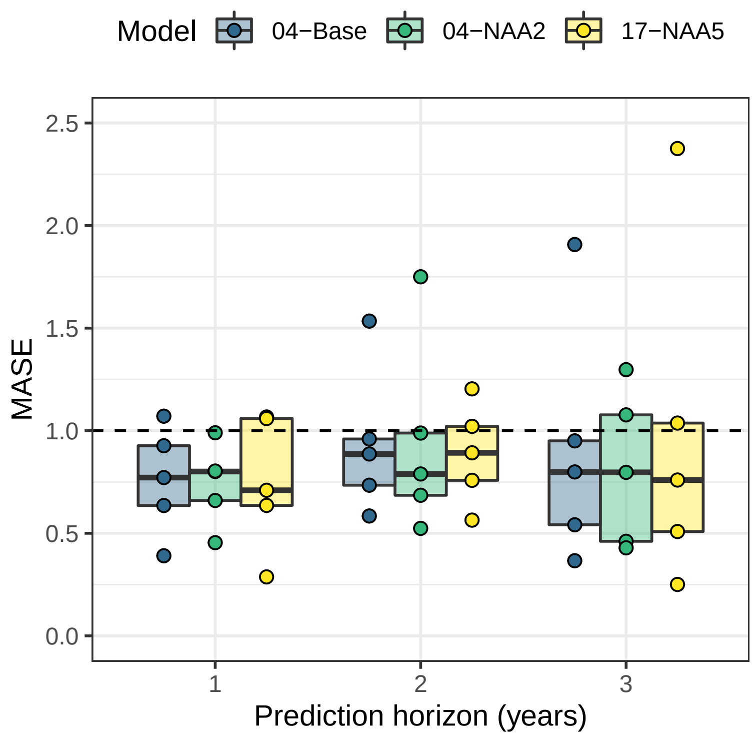

Figure 7: Hindcast performance of the three WHAM models, as measured by mean absolute scaled error (MASE). Points are the average MASE of a given model predicting each index at the specified time horizon. We performed the analysis only for the 5 indices with data in the last 3 years (Index 1 stops in 2008). MASE < 1 means that the model is better than the naive/baseline forecast, and MASE $= 0.5$ means that model forecasts are 2x as accurate as naive/baseline.

and 04-NAA2 had poor convergence rates because the index selectivity parameters for older ages were poorly estimated (SE > 3 on logit-scale). None of the models have major retrospective patterns or trends in Index-1 residuals.

17-NAA5 is also preferred on first principles for two reasons. First, the logistic normal distribution used for the age compositions is self-weighting and allows more general correlation structure than the multinomial and it has outperformed the multinomial in simulation studies (Fisch et al., 2021; Francis, 2014). Second, treating recruitment as an AR(1) process is parsimonious given the decrease in butterfish recruitment over time, and the AR(1) propagates the expectation of less than average recruitment into short-term projections in an objective fashion.

Table 3: Summary of diagnostics for the three WHAM butterfish models. Conv. = convergence rate of simulation self-tests. 'Trend Index-1' refers to trend in Index-1 residuals (Fig. 1).

					Mohn's ρ			MASE (median)		
Model	NAA random effects	Age comp	Trend Index-1	Conv.	R	SSB	F	1y	2y	3y
04-Base	—	Multinomial	None	8%	-0.07	0.01	-0.11	0.77	0.89	0.80
04-NAA2	Recruits, AR1	Multinomial	Mild	40%	0.00	0.00	-0.08	0.80	0.79	0.80
17-NAA5	All NAA, AR1	Logistic-normal	Mild	95%	0.09	0.09	0.01	0.71	0.89	0.76

3.9 Reference points and status determination

In 2019, the butterfish stock was not overfished ($B_{2019}/B_{50\%} > 1$) or experiencing overfishing ($F_{2019}/F_{50\%} < 1$) in all models (Table 4).

Table 4: Reference points and stock status in the terminal assessment year (2019) with 95% confidence intervals given in parentheses. $B_{50\%}$ = spawning stock biomass at 50% of reproductive potential, i.e. spawning potential ratio (B_0). Biomass units are metric tons (mt).

Model	$F_{50\%}$	$B_{50\%}$	$F_{2019}/F_{50\%}$	$B_{2019}/B_{50\%}$
04-Base	4.92	29360	0.06 (0.03–0.10)	1.94 (1.20–3.14)
04-NAA2	4.74	32680	0.05 (0.03–0.10)	1.73 (0.96–3.11)
17-NAA5	6.62	37318	0.04 (0.02–0.08)	2.08 (1.20–3.63)

The three WHAM models estimated similar trends in SSB, F, and recruitment as ASAP (Fig. 8). The main difference is that the state-space models, 04-NAA2 and 17-NAA5, estimated less of a decrease in recruitment and SSB over the time-series, i.e. lower recruitment and SSB in early years and higher recruitment and SSB in later years. Recruitment in the full state-space model, 17-NAA5, was notably higher in 2008-2019 (Fig. 8), which corresponds to a period of negative survival deviations for fish aged 1+ (Fig. 3).

All models estimated that the stock has never been overfished or experienced overfishing ($B/B_{50\%} > 1$ and $F/F_{50\%} < 1$ in all years, Fig. 9).

3.10 Projections

We show 3-year projections of recruitment and SSB under 3 alternative F scenarios: $F = 0$ (Fig. 10), $F = F_{2019}$ (Fig. 11), and $F = F_{50\%}$ (Fig. 12). 17-NAA5 estimates a larger population size (higher SSB) than the other two models, even with similar F and the same M, because survival deviations for ages 1+ were estimated to be positive (Figs. 3 and 10-12).

Note the effect of treating projected recruitment as a continuation of the AR(1) process (or not, as in 04-Base). Recruitment in 04-Base jumps immediately to the 2011-2019 average but recruitment in 04-NAA2 and 17-NAA5 gradually approach average recruitment (Figs. 10-12).

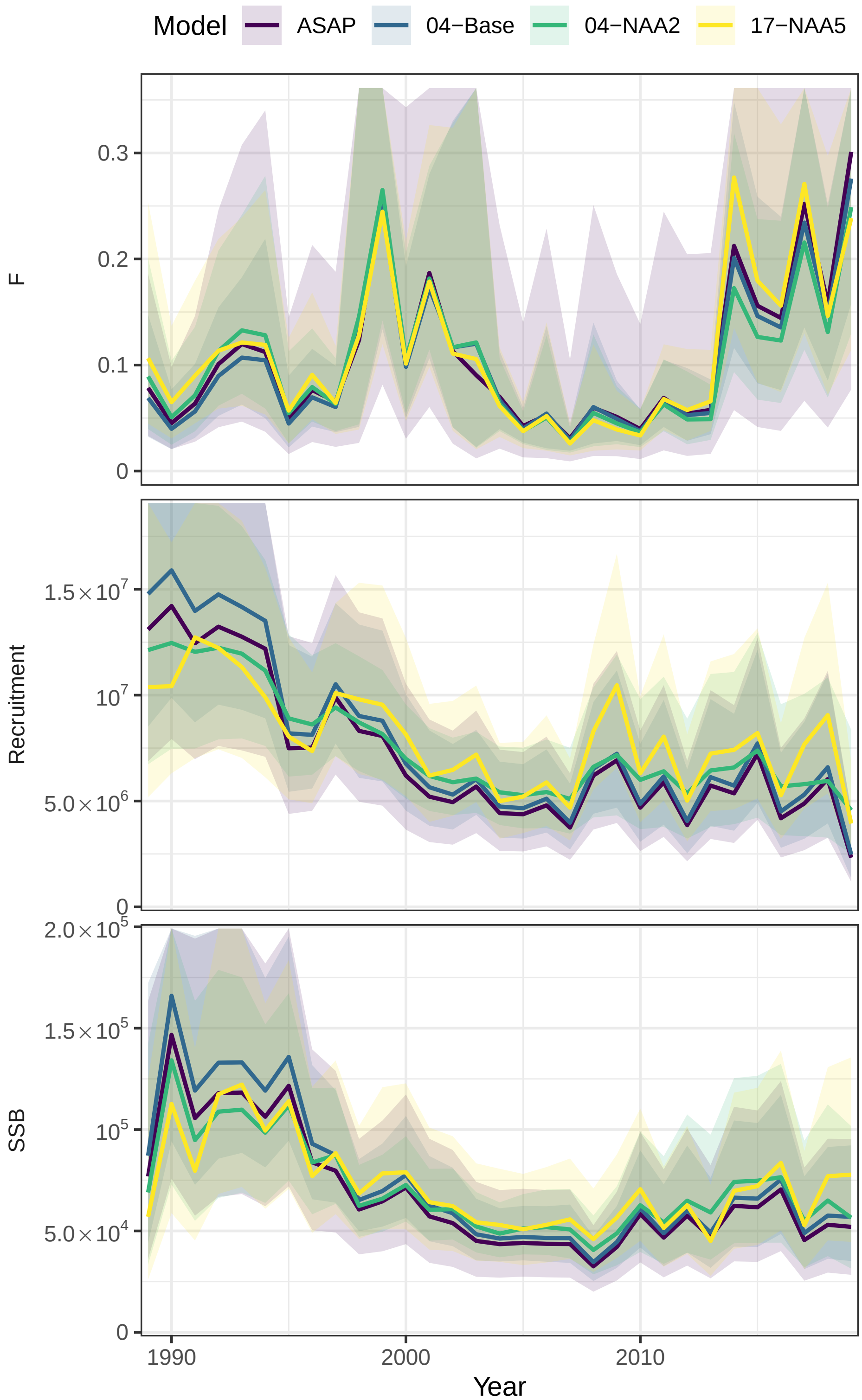

Figure 8: Spawning stock biomass (SSB), fishing mortality (F), and recruitment estimated by ASAP and the three final WHAM models.

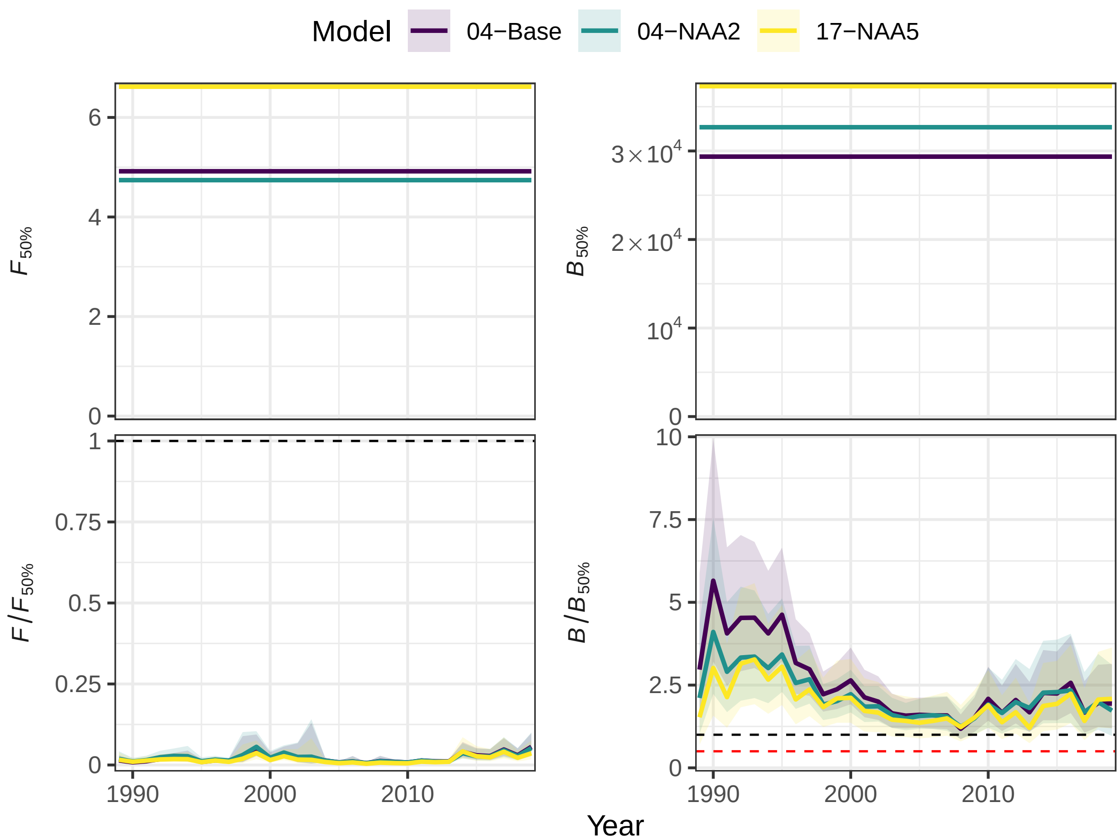

Figure 9: Butterfish reference points and stock status. $B_{50\%}$ = spawning stock biomass at 50% of reproductive potential, i.e. spawning potential ratio (SPR, B_0), and $F_{50\%}$ = fishing mortality at 50% SPR. Black horizontal dashed lines indicate $F/F_{50\%} = 1$ and $B/B_{50\%} = 1$. Red dashed line indicates $B/B_{50\%} = 0.5$.

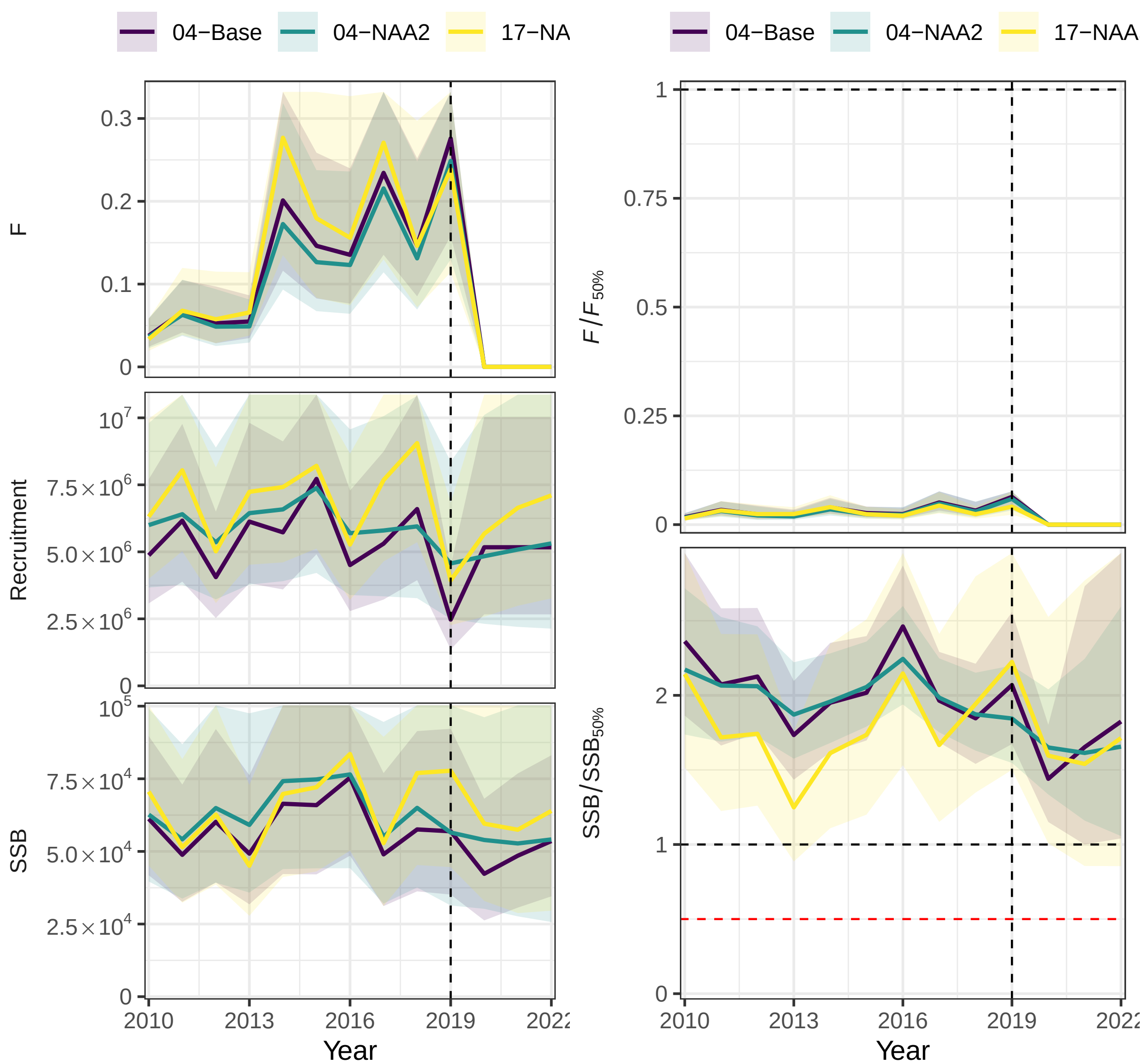

Figure 10: Spawning stock biomass (SSB), fishing mortality (F), and recruitment estimated by WHAM models in the final 10 assessment years (2010-2019, left of vertical dashed line) and projection period (2020-2022, right of vertical dashed line) under the $F = 0$ projection scenario. Black horizontal dashed lines indicate $F/F_{50\%} = 1$ and $B/B_{50\%} = 1$. Red dashed line indicates $B/B_{50\%} = 0.5$.

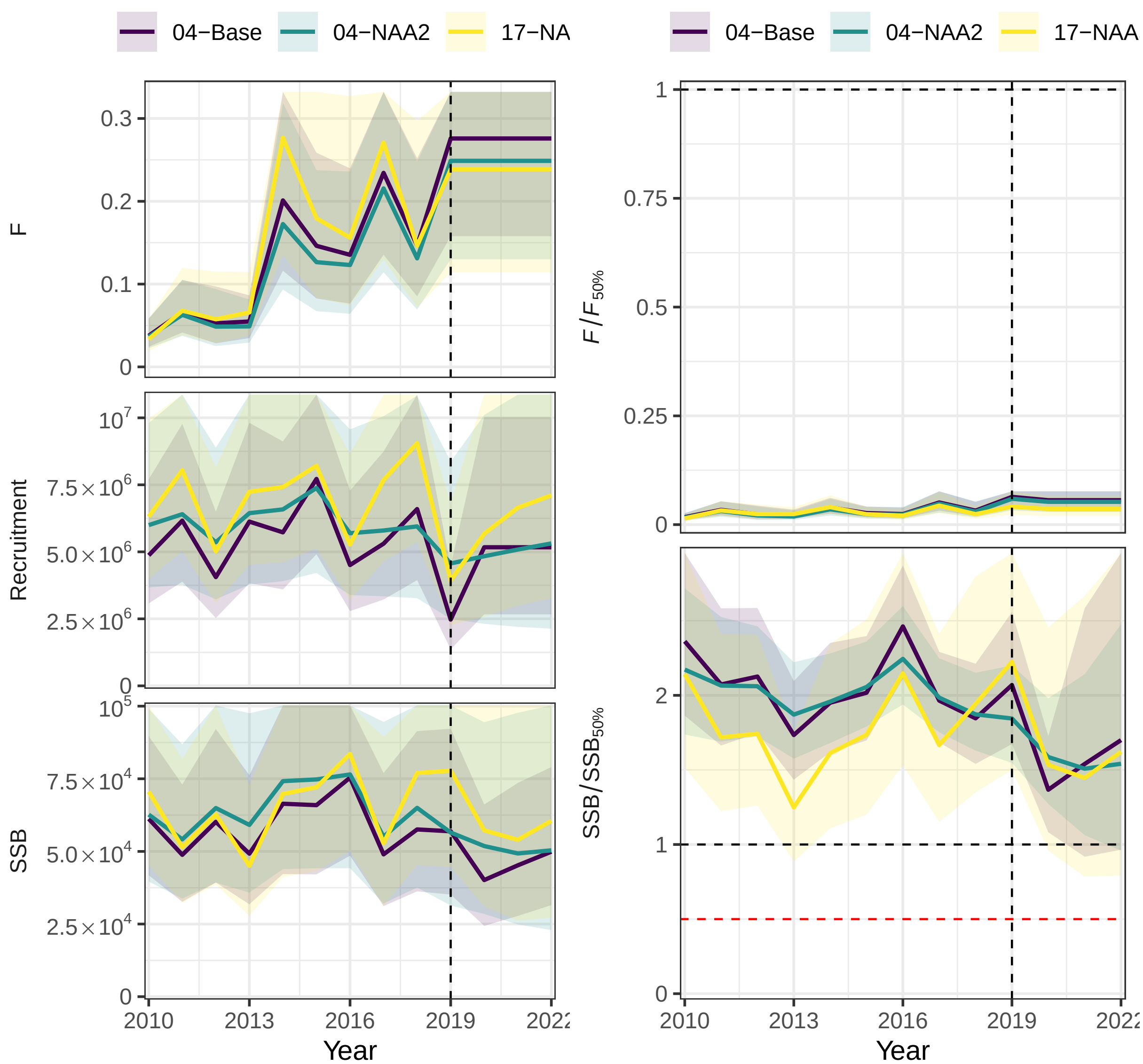

Figure 11: Spawning stock biomass (SSB), fishing mortality (F), and recruitment estimated by WHAM models in the final 10 assessment years (2010-2019, left of vertical dashed line) and projection period (2020-2022, right of vertical dashed line) under the $F = F_{2019}$ projection scenario. Black horizontal dashed lines indicate $F/F_{50\%} = 1$ and $B/B_{50\%} = 1$. Red dashed line indicates $B/B_{50\%} = 0.5$.

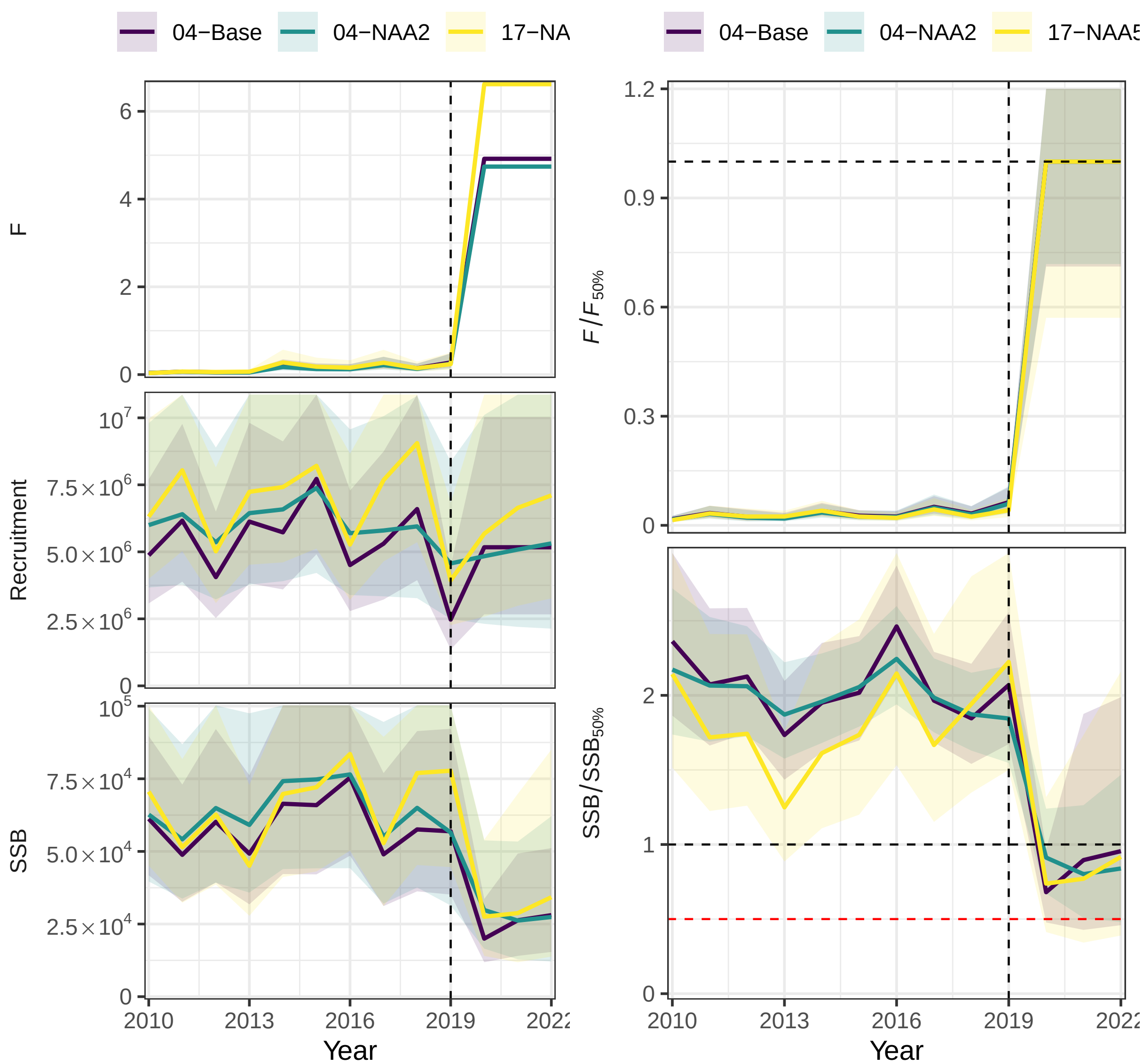

Figure 12: Spawning stock biomass (SSB), fishing mortality (F), and recruitment estimated by WHAM models in the final 10 assessment years (2010-2019, left of vertical dashed line) and projection period (2020-2022, right of vertical dashed line) under the $F = F_{50\%SPR}$ projection scenario. Black horizontal dashed lines indicate $F/F_{50\%} = 1$ and $B/B_{50\%} = 1$. Red dashed line indicates $B/B_{50\%} = 0.5$.

References

Aeberhard, W.H., Mills Flemming, J., Nielsen, A., 2018. Review of State-Space Models for Fisheries Science. Annu. Rev. Stat. Appl. 5, 215–235. https://doi.org/10.1146/annurev-statistics-031017-100427

Berg, C.W., Nielsen, A., 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES J Mar Sci 73, 1788–1797. https://doi.org/10.1093/icesjms/fsw046

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., Methot, R.D., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fisheries Research 240, 105959. https://doi.org/10.1016/j.fishres.2021.105959

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. Fisheries Research 243, 106069. https://doi.org/10.1016/j.fishres.2021.106069

Francis, R.I.C.C., 2014. Replacing the multinomial in stock assessment models: A first step. Fisheries Research 151, 70–84. https://doi.org/10.1016/j.fishres.2013.12.015

ICES, 2020. Workshop on the review and future of state space stock assessment models in ICES (WKRFSAM). ICES Scientific Reports 2, 23p. https://doi.org/10.17895/ices.pub.6004

Kell, L.T., Sharma, R., Kitakado, T., Winker, H., Mosqueira, I., Cardinale, M., Fu, D., 2021. Validation of stock assessment methods: Is it me or my model talking? ICES Journal of Marine Science fsab104. https://doi.org/10.1093/icesjms/fsab104

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., Bell, B.M., 2016. TMB: Automatic differentiation and Laplace approximation. Journal of Statistical Software 70, 1–21. https://doi.org/10.18637/jss.v070.i05

Legault, C.M., Bell, R.J., Brooks, E.N., Cournane, J., Deroba, J.J., Fay, G., Jones, A., Langan, J., Miller, T.J., Muffley, B., Wiedenmann, J., 2021. Data Rich but Model Resistant: An Evaluation of Data-Limited Methods to Manage Fisheries with Failed Age-Based Stock Assessments. NOAA Fisheries, Northeast Fisheries Science Center.

Legault, C.M., Restrepo, V.R., 1998. A Flexible Forward Age-Structured Assessment Program (No. 49).

Methot, R.D., Taylor, I.G., 2011. Adjusting for bias due to variability of estimated recruitments in fishery assessment models. Can. J. Fish. Aquat. Sci. 68, 1744–1760. https://doi.org/10.1139/f2011-092

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

Miller, T.J., Hyun, S.-Y., 2018. Evaluating evidence for alternative natural mortality and process error assumptions using a state-space, age-structured assessment model. Can. J. Fish. Aquat. Sci. 75, 691–703. https://doi.org/10.1139/cjfas-2017-0035

Miller, T.J., Legault, C.M., 2015. Technical details for ASAP version 4 (No. Ref Doc. 15-17). US Dept Commer, Northeast Fish Sci Cent.

Miller, T.J., Stock, B.C., 2020. The Woods Hole Assessment Model (WHAM).

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research 158, 96–101. https://doi.org/10.1016/j.fishres.2014.01.014

Stock, B.C., Miller, T.J., 2021. The Woods Hole Assessment Model (WHAM): A general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates. Fisheries Research 240, 105967. https://doi.org/10.1016/j.fishres.2021.105967

Stock, B.C., Xu, H., Miller, T.J., Thorson, J.T., Nye, J.A., 2021. Implementing two-dimensional autocorrelation in either survival or natural mortality improves a state-space assessment model for Southern New England-Mid Atlantic yellowtail flounder. Fisheries Research 237, 105873. https://doi.org/10.1016/j.fishres.2021.105873

Thorson, J.T., 2019. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fisheries Research 217, 133–139. https://doi.org/10.1016/j.fishres.2018.02.005

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics 24, 317–339. https://doi.org/10.1007/s10651-017-0372-4

Xu, H., Thorson, J.T., Methot, R.D., 2020. Comparing the performance of three data weighting methods when allowing for time-varying selectivity. Can. J. Fish. Aquat. Sci. 77, 247–263. https://doi.org/10.1139/cjfas-2019-0107

Table 5: WHAM runs with description and comments. Bold rows indicate models presented in detail for consideration.

Run	Description	Comments
1	As specified in asap3 RUN 36, except index 1 q is freely estimated. Try 5 standard NAA models.	None converge.
2	As 01, except fix index 1 q at the initial value from asap3 RUN 36 (0.21). Try 5 standard NAA models.	Only NAA2 converges. Others would if index 6 (neamap-spring) selAA-3 is fixed at 1 (estimated 15 on logit scale). Also noted selAA pars are initialized at bound (1) instead of middle of range (0.5).
3	As 02, except fix index 6 (neamap-spring) sel-at-age-3 at 1. Initialize selAA pars at 0.5.	Base, NAA1, and NAA2 converge with max gradient < 3e-12. Full state-space models (NAA3 and NAA4) have estimation problems for sigma-a (goes to 0 with NaN or high SE).
4	**As 03, except fix index 1 q at the estimated value from asap3 RUN 36, 0.197517.**	**NLL is 0.1 lower with q1 = 0.1975 than q1 = 0.21. 04-Base and 04-NAA2 worth considering.**
5	As 03, except estimate mean M.	M is estimable, lower than fixed in ASAP (1.278). Mean with 95% CI: Base 1.00 (0.65-1.55), NAA1 0.92 (0.59-1.44), NAA2 0.95 (0.59-1.50).
6	Try to estimate q1 if the extra selectivity parameter is fixed and M is estimated.	Fail.
7	Likelihood profile over q1.	Fail, wants to go to -Inf.
8	As 05. Don't fit to catch paa in years without data, 1998-2013. Pred catch paa in those years uses selectivity shared with 89-97.	
9	As 08 except without estimating M.	
10	As 03. Use Dirichlet-multinomial age composition likelihood instead of multinomial.	
11	As 03. Use logistic-normal age composition likelihood instead of multinomial.	
12	As 11 but try to estimate q1 again.	Fail.
13	As 11 but estimate M.	All converge. Not supported by AIC for state-space models but is for NAA1.
14	As 11 but estimate Beverton-Holt stock recruitment.	All converge. NAA3 best by AIC.
15	As 14 but use multinomial age comp likelihood.	Fail.
16	As 14 but try initializing FMSY higher.	Still only get FMSY estimates for half of years.
17	**As 11 but fix q1 at value estimated in asap, 0.1975**	**17-Base and 17-NAA2 look good, 17-NAA5 has best AIC and prediction skill.**
18	As 17 but estimate M.	
19	As 17 (logistic normal age comp for fleet, with lots of age data) but use multinomial for indices age comp.	No models with rho < 0.05 and no index 1 trend.
20	Like 19 but swapped, multinomial fleet / logistic normal indices.	No models with rho < 0.05 and no index 1 trend.
21	As 17 but pool zeros instead of treat as missing.	Mohn's rho higher.
22	Attempt to put AR1 random effects on selectivity	Convergence issues.
23	Multinomial age comp, AR1 random effects on selectivity.	Convergence issues.
24	As 23 but with logistic normal age comp.	Issues with index 1 residual trend and retros.
25	One time-varying sel block for fishery, try age-specific and logistic with 2D AR1 random effects.	25-NAA2-selAA looks ok but diagnostics are not as good as 17-NAA5.