Assessment Model Development and Stock Projections for the Southern New England / Mid-Atlantic Yellowtail Flounder Stock

Working Paper for ToRs 4-6

Cameron Hodgdon[1]

[1]Northeast Fisheries Science Center, Woods Hole, MA

1.0 Abstract

An age-structured assessment model (Woods Hole Assessment Model (WHAM)) was developed for the southern New England / mid-Atlantic (SNEMA) yellowtail flounder stock for the period 1973-2022. The model used aggregate data from commercial landings and discards, three fishery-independent surveys: the NEFSC Spring, NEFSC Fall, and NEFSC Winter bottom trawl surveys, and an index of the northern extent of the Gulf Stream (GSI) to inform recruitment trends. Overall, the SNEMA stock of yellowtail flounder is doing poorly in comparison to the historical period, with record low recruitment and SSB in the recent period. Projections of the stock while maintaining a fishing pressure equal to that of the F_{MSY} proxy, $F_{40\%}$, reveal both short and long-term increases to landings, SSB, and R. However, projections are dependent on the methodology used to project the GSI.

2.0 Introduction

Previously, the stock of yellowtail flounder in the southern New England / mid-Atlantic (SNEMA) region (Figure 2.0.1) was assessed using the age-structured assessment program, ASAP (NEFSC 2012; 2022). This assessment had very strong retrospective patterns due in part to low or missing survey information in the recent time series (NEFSC 2022). Additionally, the long-term projections for this stock were uncertain (NEFSC 2022) and possibly strongly correlated with environmental effects (Xu et al. 2017; Stock and Miller 2021; du Pontavice et al. 2022). In an attempt to mitigate these issues, herein we develop an aged-structured stock assessment model for the SNEMA stock of yellowtail flounder using the Woods Hole Assessment Model (WHAM). WHAM is a state-space, age structured model capable of including environmental effects on population processes, estimating both observation and process error, and propagating random effects (Stock and Miller 2021). In this paper, model inputs, model selection process, and model results are summarized. Additionally, biological reference point determination and projection methodologies are discussed.

3.0 Data

3.1 Fishery Data

The SNEMA stock of yellowtail flounder once supported high catches (Figure 3.1.1). Throughout the last several decades, catches have declined and today, the SNEMA stock is a bycatch fishery (Figures 3.1.2 and 3.1.3), where most of the yellowtail landed are with bottom trawls (Figure 3.1.4). For details of landings and discards and a discussion of these data by statistical areas and gear types, see ToR 2 in the Yellowtail Flounder Research Track Working Group Report.

WHAM combines commercial landings and discards into a single fleet (Figure 3.1.1). Age composition data is available for this combined fleet for most of the time series with the exception of the last few years (Figure 3.1.5), where biological sample collection was constricted by immensely low catches. Age compositions of the combined fleet were historically centered on ages 2-4, but in the recent period have shifted to ages 4-6. It is important to note here, however, that the derivation of age compositions in the recent period is challenging due to the lack of biological samples. For a full breakdown of this data, see the Yellowtail Flounder Research Track Working Group Report ToR 2 chapter.

3.2 Survey Data

Of the available fishery-independent survey datasets available for the SNEMA stock, five were originally considered for inclusion in the WHAM framework. These five surveys were the NEFSC spring, NEFSC fall, and NEFSC winter bottom trawl surveys as well as two larval indices based on Richardson et al. (2010). These five surveys were used in previous assessments and a full breakdown of the surveys themselves can be found in the Yellowtail Flounder Research Track Working Group Report ToR 3 chapter. However, due to data availability, the two larval indices were not included in the candidate model.

4.0 Assessment Model Development

4.1 Initial Parameterization

Due to WHAM's ability to emulate ASAP, the initial setup was identical to the setup of the last management track assessment of SNEMA yellowtail flounder as discussed in NEFSC (2022). This setup included the number of age classes (1-6+) and CV on the combined fleet landings (0.05). From here, over 400 model variations were tested and compared that considered alternative assumptions on fleet and survey selectivities, recruitment, age compositions, life history parameters, and environmental effects. Each of these decisions are documented below, but as should be noted, the below decisions were not conducted in a linear fashion, but rather past decisions were revisited when necessary due to model changes from latter decisions.

4.2 Fishery Selectivity

Initial fleet selectivity was assumed to follow an asymptotically increasing function (logistic selectivity) over ages 1-6+, with selectivity near zero at age-1 and at 1 at older ages. Age-specific selectivity inputs were also compared, but this model configuration did not pass second-order convergence criteria. Random effects on logistic fleet selectivity in the form of 1) independent and identically distributed (iid), 2) autoregressive (ar1) process between logistic equation parameters, 3) autoregressive (ar1) process across years, and 4) autoregressive process across both parameters and years (2dar1) were tested in the WHAM configuration. Of these, the iid random effects led to an overall model with better fits and retrospective patterns than the basecase with no random effects. Thus, logistic fleet selectivity with iid random effects was carried to the candidate model configuration (Figures 4.2.1 and 5.2.4).

Fleet selectivity was also blocked into a 1973-1995 selectivity and a 1996-2022 selectivity. This was done to test the hypothesis that the time around 1995 may have had a significant impact on selectivity, as this period saw the introduction of Vessel Trip Reporting (VTRs) and mandatory logbook reporting. This configuration within WHAM could not converge, however, and so was abandoned. Thus, fleet selectivity in the candidate model remained a single block representing selectivity 1973-2022 with iid random effects allowing the data to guide changes in selectivity across years.

4.3 Fishery Age Composition

The default age composition structure in WHAM is multinomial (to match the structure used in ASAP). Aside from this, age composition structures of dirichlet, dirichlet multinomial, logistic normal, and multivariate tweedie were tested. For details on these different structures, see Stock and Miller (2021). Additionally, dirichlet and logistic normal age compositions have the ability to either treat the zero observations in the data as missing (miss0) or to pool zero observations with adjacent (non-zero) age classes (pool0). Logistic normal-miss0 could also have an added autoregressive process (ar1) to relate neighboring age classes (logistic normal-ar1-miss0). This led to a total of ten different age structures tested: multinomial, dirichlet multinomial, dirichlet-miss0, dirichlet-pool0, logistic normal-miss0, logistic normal-pool0, logistic normal-ar1-miss0, and multivariate tweedie.

Multivariate tweedie was abandoned due to very large computation times, making a model selection process with multivariate tweedie not feasible. Logistic normal-miss0 led to relatively good residuals, fits, and retrospectives, but the addition of an ar1 process in this structure led to even better diagnostics. Thus, the age composition structure which was put forward in the candidate model for the fleet was logistic normal-ar1-miss0 (i.e. a logistic normal age composition that treats zero data as missing and that includes an autoregressive (ar1) process to relate neighboring age classes).

4.4 Survey Selectivity

Similar to what was done for the combined fleet, each of the five surveys in the initial model were run with logistic and age-specific selectivities. Logistic selectivities led to better fits and retrospective patterns for the NEFSC fall, spring, and winter bottom trawl indices and age-specific selectivities led to better fits and retrospective patterns for the two larval indices. Additionally, both independent and identically distributed (iid) random effects and autoregressive (ar1) across years random effects were tested on the three NEFSC survey selectivities. Neither of these random effects led to a better fitting model. The two larval indices were not similarly tested because in order to obtain model convergence, all ages needed to be fixed for the selectivities (i.e. there was no fitting done for the larval index selectivities; selectivity was input). This was one of the reasons for larval index exclusion, which will be discussed further in section 4.7.

In 2009, the NEFSC switched from the Albatross to the Bigelow for the vessel which was used to carry out the bottom trawl surveys (additionally, during this transition, the gear and tow time changed). To test the effect of the vessel switch on SNEMA yellowtail data, a model run was conducted where selectivity was blocked. The first selectivity block represented Albatross selectivity 1973-2008 and the second block represented Bigelow selectivity 2009-2022. Differences were minimal; the inclusion of this blocking effect did not lead to better retrospective patterns or residuals. Thus, this blocking was not included in the candidate model and a single selectivity 1973-2022 was used with iid random effects.

Late in the model selection process, an update to WHAM was made to more appropriately deal with random effects on recruitment when an environmental covariate is also used (discussed in section 4.9), and after this update, some previous candidate model decisions were re-visited. A change that came from this revisitation was the changing of the NEFSC fall bottom trawl survey selectivity from a logistic function to an age-specific selectivity. Thus, in the candidate model, NEFSC spring and winter bottom trawl surveys have a logistic selectivity, but the NEFSC fall bottom trawl survey uses an age-specific selectivity that very rapidly approaches 1 (starting values are 0.5 for age-1 and 1 for ages 2-6+).

4.5 Survey Age Compositions

Similar to what was done for the age composition of the aggregate fleet, multiple structures were tested for each of the five survey-based indices. However, multivariate tweedie was not explored for the reasons noted in section 4.3. Similar to findings for the aggregate fleet, logistic normal-ar1-miss0 was found to be the best age composition structure for all five indices: the NEFSC spring, fall, and winter bottom trawl indices and the two larval indices.

4.6 Survey Catchability

Random effects in the form of both independent and identically distributed (iid) and autoregressive (ar1) across years were examined for all five surveys. Random effects on any of

the three NEFSC surveys did not make it into the candidate model as the addition of them did not improve model fit or retrospective patterns. However, the larval surveys needed catchability random effects in the form of ar1. The exclusion of these random effects on larval index catchability led to extremely unsatisfactory residuals and retrospective patterns.

Effects of bottom temperature on catchability were also examined for the three NEFSC surveys (this was done after the decision to remove the larval indices was made - see section 4.7). Bottom temperature was subset to the season the respective survey operates (e.g. spring bottom temperature was used for the NEFSC spring survey) and the covariate was modeled as both a random walk and an autoregressive (ar1) process. None of these effects (all were done with a linear link) of bottom temperature on the NEFSC bottom trawl indices improved residuals or retrospective patterns and were so not carried over to the candidate model.

4.7 Survey Inclusion/Exclusion

As stated above, five fishery-independent surveys were originally considered for inclusion in the WHAM framework. These five surveys were the NEFSC spring, NEFSC fall, and NEFSC winter bottom trawl surveys as well as two larval indices based on Richardson et al. (2010). After appropriate selectivities and catchabilities had been determined for each of these surveys (see above), a full "leave-one-out" analysis was conducted where one survey at a time was dropped from the WHAM framework and the model re-run. This was also done for every combination of two surveys, three surveys, and four surveys. These 32 runs brought into question the validity of the larval indices again, which usually led to worse retrospective patterns when they were included in the model run. It was decided to drop the larval indices from the candidate model for the following reasons:

1) Selectivity of the larval indices had to be fixed at all ages, otherwise there was no model convergence, pointing to the extreme rigidity/sensitivity of the index.
2) Catchability of the larval indices was not realistic due to the large uncertainty and extreme inter-annual variation (Figure 4.7.1) and the addition of random effects on catchability was necessary for model convergence, again pointing to the instability and sensitivity of the indices.
3) The last few years of data had extremely low sample sizes. As is discussed in more detail in the Yellowtail Flounder Research Track Working Group Report ToR 3 chapter, this index takes into account larval size in the index estimation, which can exacerbate issues of low sample sizes. For example, two individual larval yellowtail flounder were caught in this survey in 2021 and two were caught in 2022. However, because the two 2021 individuals were larger than the two 2022 individuals, the index showed a decline of ~93% between 2021 and 2022. The noise at these low sample sizes cannot be reflective of true population trends and this problem would persist if the survey continues to catch very few flounder, exacerbating this issue in subsequent assessments.

4.8 Natural Mortality Assumptions

Natural mortality (M) was assumed to be constant across ages and years. A value of M = 0.5 was used based on a longevity estimate (for details, see the Yellowtail Flounder Research Track Working Group Report ToR 1 chapter). Random effects in the form of 1) independent and identically distributed (iid), 2) autoregressive (ar1) process across ages, 3) autoregressive (ar1) process across years, and 4) autoregressive process across both ages and years (2dar1) were tested in the WHAM configuration, but many did not converge (and the few that did had worse fits and worse retrospective patterns). Age-specific estimates of M were also tested with the random effects listed above and led to similar results (most configurations did not converge and those that did have worse fits and retrospective patterns when compared to the basecase of M = 0.5 across ages and years).

Due to the results for SNEMA yellowtail flounder from ToR 1 (see the Yellowtail Flounder Research Track Working Group Report ToR 1 chapter), effects of the Atlantic Multidecadal Oscillation, average fall bottom temperature, and average spring bottom temperature were tested on M. Effects of each of these were modeled as both linear and polynomial-2. For details on each of these effects, see Stock and Miller (2021). Each of the environmental covariates were also modeled as either a random walk or an autoregressive (ar1) process. None of these 64 combinations led to a converged model with better fits or retrospective patterns than the basecase assumption of M = 0.5 across all ages and years. This assumption (M = 0.5) was put forth in the candidate model.

4.9 Recruitment Assumptions

Prior to the testing of environmental covariate effects on recruitment, a full analysis was done on recruitment assumptions. Recruitment was modeled as random around a mean, random walk, through a Beverton-Holt stock/recruit relationship, and through a Ricker stock/recruit relationship. In the absence of any environmental effects on recruitment, both Beverton-Holt and Ricker stock/recruit relationships improved overall model diagnostics (with Beverton-Holt providing slightly better diagnostics than Ricker).

Based on analyses outlined in the Yellowtail Flounder Research Track Working Group Report ToR 1 chapter, effects of Atlantic Multidecadal Oscillation, North Atlantic Oscillation, spring bottom temperature, the Cold Pool Index, the Gulf Stream Index, and the spring Gulf Stream Index were all tested for inclusion in the WHAM framework as a covariate which influenced recruitment trends.

For all of these tests, recruitment was decoupled from numbers-at-age (NAA), meaning that age-1 individuals could have trends independent over time of the trends experienced by ages 2-6+. Additionally, for all of these tests, the Beverton-Holt stock-recruit relationship was removed (and recruitment was modeled as random around a mean) and replaced by the effect of the environmental covariate. This was done because of the inability to develop a converged

model that incorporated both a stock-recruit relationship and the effect of a covariate. This assumption can be thought of as replacing the effect of the spawning biomass on recruitment with the effect of the environmental covariate on recruitment.

Each covariate tested was modeled as both a random walk process and an autoregressive (ar1) process and could affect recruitment through a "controlling", "limiting", or "masking" effect (Stock and Miller 2021) that was lagged by one year under the assumption that recruitment would be most influenced by environmental conditions one year prior.

Of all of these combinations, the "controlling" effects of spring bottom temperature and the Gulf Stream Index both improved residuals and retrospective patterns over the basecase with no environmental effects (to note, "limiting" and "masking" effects almost never converged for all covariates). Of these two effects, GSI had better diagnostics and was chosen to be included in the candidate model. GSI in the model performed well when modeled with a random walk process and an ar1 process. The ar1 process was chosen as the preferred method because this allowed for more accurate forecasts of GSI.

4.10 Numbers-at-Age

Numbers-at-age in the WHAM framework can also include random effects. These effects can be independent and identically distributed (iid), autoregressive (ar1) across years, autoregressive (ar1) across ages, and autoregressive across both ages and years (2dar1). Autoregressive processes across both ages and years (2dar1) improved overall model fit and retrospective patterns, leading to inclusion in the candidate model. Due to recruitment being decoupled from ages 2-6+ (see section 4.9), the autoregressive (ar1) process across ages only applied to ages 2-6+ and not between age-1 and age-2.

4.11 Candidate Model Setup

All candidate model settings based on the above criteria can be found in Table 4.11.1. The model (nicknamed m164_GSI) uses logistic fleet selectivity with iid random effects, logistic NEFSC spring and winter selectivities, and age-specific NEFSC fall selectivity. The age composition of the aggregate fleet and all surveys is modeled as logistic normal-ar1-miss0 and NAA have 2dar1 random effects. Recruitment (age-1) is decoupled from ages 2-6+ (but still have an autoregressive process across years) and trends are informed by GSI (modeled via ar1) lagged one year via a "controlling" process.

5.0 Model Results

5.1 Candidate Model Diagnostics

The model put forth as the candidate passed both first-order convergence criteria (max gradient 5.11e-11) and second-order convergence criteria (Hessian matrix was invertible). The jitter analysis confirmed that the model converged on a global solution and was very stable with a

convergence rate of 93% (Figure 5.1.1).

Given a low assumed CV for the aggregate fleet (commercial landings and discards) of 0.05, the model fit the data very well with minimal patterning (Figure 5.1.2). Fits to the NEFSC Spring, NEFSC Fall, and NEFSC Winter bottom trawl surveys were also relatively good, with little residual patterning over the time series (Figures 5.1.3, 5.1.4, and 5.1.5) and only a few outlier years across all three surveys in which the fits were outside the confidence bounds. Fits to GSI were very good, with little residual patterning and no years with which the confidence bounds of the GSI did not overlap with the confidence bounds of the prediction (Figure 5.1.6).

OSA residual diagnostics for the aggregate fleet, all three survey indices (NEFSC Spring, NEFSC Fall, and NEFSC Winter), and GSI were relatively normally distributed (Figures 5.1.7, 5.1.8, 5.1.9, 5.1.10, and 5.1.11). Additionally, OSA residual diagnostics for the age compositions of the aggregate fleet and all three survey indices were relatively normally distributed (Figures 5.1.12, 5.1.13, 5.1.14, and 5.1.15). Bubble plots of OSA quantile residuals are also presented below (Figures 5.1.16, 5.1.17, 5.1.18, and 5.1.19) and did not show any major diagnostic issues.

Retrospective analysis revealed that the candidate model had a minor tendency to overestimate SSB (Mohn's ρ = 0.145; Figures 5.1.20 and 5.1.21) and a minor tendency to underestimate F (Mohn's ρ = -0.014; Figures 5.1.22 and 5.1.23). For both SSB and F, 2020 was the year with the largest peel, being the year without survey data due to the Covid-19 pandemic. This suggests that missing survey data in the terminal year of the assessment increases uncertainty in terminal year estimates of both SSB and F. Even though it was not an official criteria, retrospective analysis also revealed that the model may have a tendency to overestimate recruits (Mohn's ρ = -0.671; Figures 5.1.24 and 5.1.25). For SSB, F, and R, peels were both above and below the terminal assessment, further indicating the lack of a problematic retrospective pattern.

AIC was used throughout the model selection process, but candidate model AIC out of context is not meaningful and cannot be compared to all alternative models discussed in the Assessment Model Development section. Regardless, AIC for the candidate model was -1948.6.

Model performance was evaluated using self-tests. The self-tests revealed the mean percent bias for F, R, and SSB were 0.78, -16.14, and -0.8, respectively with an overall convergence rate of 90% (Figure 5.1.26).

5.2 Candidate Model Estimates

Predictions from the candidate model show that fishing mortality has, in general, decreased since the historical period (Figure 5.2.1) and is currently at F = 0.11. During this same period, SSB has decreased substantially (Figures 5.2.1 and 5.2.2) and 2020 represents the lowest SSB of the time series at 44 mt, less than 1% of the highest historical levels. These SSB estimates are also of similar magnitude to chainsweep expanded survey biomass estimates generated using NEFSC bottom trawl survey data (Figure 5.2.3; for details on chainsweep expanded survey biomass

estimates, see ToR 8 in the Yellowtail Flounder Research Track Working Group Report), especially in the most recent period, reflecting the historical lows of the stock. Also during this same period, R has decreased substantially (Figure 5.2.2) and is currently at the lowest in the time series at 495,000 individuals.

The CVs of F 1973-2022 were all between 0.2 and 0.5 (Figure 5.2.4), with the terminal year representing the most uncertainty. CVs of SSB were, in general, lower than those for F, varying between 0.25 and 0.31, with the exception of the terminal years, reaching above 0.4 (Figure 5.2.4). CVs of R were the largest, varying between 0.35 and 0.7, with years 2020-2022 representing the most uncertain period (Figure 5.2.4). Additionally presented below are estimated aggregate fleet and survey selectivities (Figure 5.2.5) as well as estimated catchabilities of the surveys (Figure 5.2.6)

Numbers-at-age (NAA) over time has been dominated by age-1 individuals, but in the recent period has seen a slight increase in the proportion of age group 6+ (Figures 5.2.7 and 5.2.8). This proportional increase in the plus group is more apparent in SSB-at-age over time, where SSB was historically composed primarily of age-2 through age-4 fish, but in the last decade, has seen a large increase in the proportion of age-6+ in the SSB (Figures 5.2.9 and 5.2.10) coinciding with record-low SSB values.

6.0 Biological Reference Points

6.1 Fishing Mortality

A spawner-per-recruit analysis was conducted to determine the F_{MSY} proxy reference point; $F_{40\%}$. For this calculation, an appropriate period for which to take an average for weight-at-age (WAA) and fleet selectivity was needed (natural mortality and maturity averages were also needed for this calculation, but they are constant over the entire time series in the candidate model and so required no analysis to determine periods over which to take an average). Fleet selectivity changes over time only via iid random effects, so efforts were focused on WAA.

A moving-window analysis was completed on WAA (Figure 6.1.1) in which weight at a given age (e.g. age-5) in year X is predicted using the average of years X-1 to X-Y, where Y varies from 2 to 10. Root mean squared error (RMSE) was estimated from summing the residuals of this analysis across all ages (1-6+) for each value of Y. This analysis was conducted a second time, but this time to predict the average of years X-3 to X (to match the time period most commonly used for “near-term” projections) from an average of years X-4 to X-(Y+3), where Y again varies from 2 to 10. The window with the lowest RMSE was the two year window (i.e. the last two years are better at predicting the next few years of WAA than the last 3+ years). The more years comprise an optimal window, the more stable WAA is over time. Thus, a moving-window for SNEMA yellowtail of two years is most likely due to the decreasing trend in WAA seen over the recent two decades for older ages (Figure 6.1.2).

Thus, two year averages of WAA and fleet selectivity were used in the calculation of the F_{MSY} proxy reference point; $F_{40\%}$ (natural mortality and maturity are also used, but are constant over time and thus do not change when taking an average over different years). This overfishing definition of $F_{40\%}$ represents the fishing mortality rate that resulted in 40% of the unfished spawning potential and was equal to 0.73 (Table 6.1.1).

6.2 Spawning Stock Biomass

Due to the effect of GSI on R in the candidate model, long-term projections of SSB at $F_{40\%}$ = 0.73 would not necessarily have matched the traditionally estimated $SSB_{40\%}$ reference point (which uses an average historical R value in its estimation). Thus, $SSB_{40\%}$ was instead estimated via long-term projections.

The associated equilibrium spawning biomass level under long-term (100 year) projections of fishing at a rate of $F_{40\%}$ = 0.73 and an average GSI of years 2012-2022 (see section 7.0) was set as the biomass reference point; $SSB_{40\%}$. $SSB_{40\%}$ was thus equal to 126 mt (Table 6.1.1). This low value for $SSB_{40\%}$ is reflective of the assumption that historical biomass is most likely unattainable for this stock in the future and the current period possibly represents a new regime of lower overall population size. The value at which catch reaches an equilibrium in this long-term projection is 97 mt. This value represents the MSY proxy that corresponds to the SSB_{MSY} proxy and F_{MSY} proxy ($SSB_{40\%}$ and $F_{40\%}$) outlined above.

6.3 Stock Status

The SSB_{MSY} proxy reference point is based on a recent period of recruitment and should not be used to interpret historical stock status; only current conditions. Terminal year (2022) $SSB/SSB_{40\%}$ = 44/126 = 0.35 and terminal year (2022) $F/F_{40\%}$ = 0.11/0.73 = 0.15. (Figure 6.3.1; Table 6.1.1).

7.0 Projections

7.1 Projection Settings

The WHAM framework has an integrated capacity to perform projections. For the candidate model, these WHAM projections incorporate uncertainty in parameter estimates and propagate forward random effects in NAA (ar1 for age-1; 2dar1 for ages 2-6+). Using WHAM, short-term projections were conducted under an assumption of setting 2023 catch equal to that of 2022 (standard for bridge-year settings in projections) and then fishing at a pressure of $F_{40\%}$ = 0.73 in years 2024-2026. For consistency with the proposed reference points discussed in section 6.0, maturity and natural mortality remained constant in the projection period while projected WAA represented an average of the terminal two years. Again, NAA random effects (ar1 for age-1; 2dar1 for ages 2-6+) were propagated into the future period.

Recruitment dynamics in the projected period are subject to the methodology used to project the GSI. The options for projecting the GSI in the WHAM framework were 1) project at a mean of the time series or a mean of a subset of the time series, 2) project using the autoregressive (ar1) process used for GSI in the candidate model, or 3) project forward a linear-based trend which is determined by regression of the time series or a subset of the time series. Linear regression-based projection methods were not considered for the projections because these increasing trends of GSI into the future climbed outside the historically seen estimates of GSI and drove recruitment and subsequent SSB to zero in long-term projections (Figure 7.1.1). However, it is worth noting from these results, that if the north wall of the GSI does continue to move further north, it will likely have detrimental effects on SNEMA yellowtail flounder recruitment.

Recruitment is highly unlikely to reach historic levels and thus, using a mean of the entire time series of GSI was also abandoned because this would drive recruitment back to a mean of the time series. Additionally, projecting GSI using the ar1 process was abandoned for the same reason; in long-term projections, the GSI returned to its mean (Figure 7.1.2). To determine a more suitable truncated period over which to take a mean of GSI to use in projections, a changepoint analysis was performed. For each year 1974-2021, GSI was split into two series with a break at the given year. A mean of both series was taken and absolute values of residuals were summed across both series to determine a total error associated with the break at that given year. The year that gave way to the lowest error was determined to represent the changepoint: 2012 (Figure 7.1.3). This does not necessarily equate to a regime shift at 2012 per se, just a recent period of GSI that is most unlike the historic series to better inform a shorter average recruitment window for realistic reference points and projections. Thus, GSI was projected forward at a mean of its 2012-2022 series (Figure 7.1.4). This kept recruitment numbers and subsequent SSB numbers relatively, and realistically, low, even in long-term projections. A long-term projection using a 2012-2022 mean GSI (and all other settings discussed above) was used to determine $SSB_{40\%}$, the equilibrium level of projected SSB (see section 6.2).

All settings for short-term projections can be found in Table 7.1.1. The assessment and projections will need to be updated with 2023-2024 data in the 2025 management track process. Thus, the projections presented below in section 7.2 should be considered provisional and not interpreted for use in management.

7.2 Projection Results

Short-term projections under the settings described above for the candidate model show increases of catch, SSB, and recruitment, but still well below historical levels (Table 7.2.1; Figures 7.2.1 and 7.2.2). Long-term projections under the settings described reveal similar trends (Figure 7.2.3 and 7.2.4). Recruitment, being very much driven by GSI (Figure 7.2.5), quickly reaches its equilibrium of 1,429,000 individuals by the third year of the projections, SSB reaches its equilibrium in just over twenty years in the long-term projections at 126 mt (the value also

used for $SSB_{40\%}$), and catch also reaches an equilibrium in just over a decade out at 97 mt (the MSY proxy).

8.0 Discussion

8.1 Detectability and Low Sample Sizes

The SNEMA stock of yellowtail flounder is at extraordinarily low numbers compared to historical estimates from this assessment using WHAM and past assessments using ASAP (NEFSC 2022). These low numbers are continuing to cause problems with detectability in fishery-independent surveys. The larval indices removal in this assessment perfectly exemplifies this problem. Historically, these larval indices were accurate representations of population trends for SNEMA yellowtail (Richardson et al. 2010; NEFSC 2022), but due to an extreme drop in detectability in recent years, they have become obsolete for the assessment (any trends have been overridden by noise).

If the detectability of SNEMA yellowtail flounder in the NEFSC trawl surveys declines to levels seen by the larval surveys, there is the potential that these surveys will also become obsolete for the SNEMA yellowtail assessment. WHAM needs improvements to be able to appropriately handle zero observations (for more details, see the Yellowtail Flounder Research Track Working Group Report ToR 7 chapter); presently, WHAM treats zero observations as missing. As zero observations in survey indices become more commonplace for depleted stocks such as SNEMA yellowtail, the need for this change in the WHAM framework continues to grow. Additionally, there is currently sufficient data to perform assessments of this stock in WHAM, but if low sample sizes continue to dominate in the survey indices, WHAM may lose its ability to accurately model the stock and more data-limited methods may have to be used.

8.2 Projections and the Gulf Stream Index

In the candidate model, recruitment trends are very much dependent on GSI, the two sharing a strong correlation across the series. It is interesting to note that in the best performing models that did not consider environmental covariates, there was a Beverton-Holt stock-recruit relationship to explain recruitment trends, which is also what Stock and Miller (2021) independently determined for SNEMA yellowtail. However, in this assessment, the addition of GSI and removal of the stock-recruit relationship led to more successful model runs (and the eventual candidate model). Stock and Miller (2021) found that the Cold Pool Index, not GSI, had the strongest correlation with recruitment trends. The Cold Pool Index was tested in this assessment process for its effects on recruitment (see section 4.9), but this index did not make it into the final candidate model. The reason for this difference between this assessment and Stock and Miller (2021) is most likely due to the addition of four years of data (2019-2022) representing the lowest recruitment sizes and also the largest values of GSI.

The candidate model uses GSI in place of a stock-recruit relationship. This is perhaps showing

that, due to the low population sizes of SNEMA yellowtail (especially in the most recent period), that recruitment sizes are less dictated by the spawning biomass and more at the mercy of environmental conditions (i.e. more variability in recruitment can be explained by GSI than can be explained by changes in the population). Consequently, projections of the candidate model are dependent on projections of GSI. Subsequent management track assessments should continuously monitor changes to GSI and ensure that WHAM estimates of future GSI agree with any WHAM-independent forecasts of the index. This will help to verify reliability in future estimates. Regardless of the methodology used to predict future trends in GSI, it is unlikely that this decision will change current perceptions of the stock.

An important note for other assessments is that the traditional method of estimating the biomass reference point $SSB_{40\%}$ should not be used if there is an environmental covariate affecting recruitment trends. This will cause a disconnect between long-term projections of SSB at $F_{40\%}$ and $SSB_{40\%}$. For these to match, $SSB_{40\%}$ should instead be estimated from the long-term projections themselves.

8.3 Missing Data and Terminal Year Estimates

From the retrospective analyses done for the candidate model and every other run completed in the model selection process (see section 4.0), a common trend was apparent: 2020 was the peel that influenced large Mohn's rho estimates (i.e. 2020 was usually the largest cause of any retrospective patterns seen). This is not surprising, as 2020 was the year in which the NEFSC bottom trawl surveys did not operate due to the Covid-19 pandemic. Thus, 2020 represents a year in the recent period in which the data for all fishery-independent indices are missing. This shows that if the terminal year of an assessment does not have fishery-independent data, there is usually going to be very large uncertainties and/or bias with terminal year estimates. If this occurs in future assessments, serious consideration should be given to the reliability of assessment results if these fishery-independent data are missing in the terminal year.

8.4 Log-normal Adjustment

WHAM has the capability to adjust estimates of recruitment and numbers-at-age through a process known as log-normal adjustment (formerly referred to as "bias correction"). For details, see Stock and Miller (2021). In the SNEMA yellowtail flounder candidate model, this log-normal adjustment is turned on. However, this decision was made in the context of model diagnostics (in this sense, log-normal adjustment was simply treated as another parameter to be changed). At the time of this writing, there is no clear guidance on the log-normal adjustment and whether it should be turned on or off (and whether this decision is stock-specific), but work is being conducted. This paragraph serves as a reminder for future management track assessments of SNEMA yellowtail to revisit this assumption of log-normal adjustment if and when more appropriate guidance on this modeling decision emerges.

References

du Pontavice, H., Miller, T. J., Stock, B. C., Chen, Z., and Saba, V. S. 2022. Ocean model-based covariates improve a marine fish stock assessment when observations are limited. *ICES Journal of Marine Science*. 79: 1259-1273.

[NEFSC] Northeast Fisheries Science Center. 2012. 54th Northeast Regional Stock Assessment Workshop (54th SAW) Assessment Report. US Dept Commer, NOAA Fisheries, Northeast Fish Sci Cent Ref Doc. 12-18; 600 p. Available online at http://nefsc.noaa.gov/publications/crd/crd1218/.

[NEFSC] Northeast Fisheries Science Center. 2022. Management Track Assessments Fall 2022. US Dept Commer, Northeast Fish Sci Cent Tech Memo. 305; 167p.+xv. Available from: National Marine Fisheries Service, 166 Water Street, Woods Hole, MA 02543-1026, or online at http://www.nefsc.noaa.gov/nefsc/publications/.

Richardson, D, Hare, J., Overholtz, W., and Johnson, D. 2010. Development of long-term larval indices for Atlantic herring (*Clupea harengus*) on the northeast US continental shelf. *ICES Journal of Marine Science*. 67(4): 617627.

Stock, B. C. and 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.

Xu, H., Miller, T. J., Hameed, S., Alade, A. L., and Nye, J. 2017. Evaluating the Utility of the Gulf Stream Index for Predicting Recruitment of Southern New England-Mid Atlantic yellowtail flounder. *Fisheries Oceanography*. DOI: 10.1111/fog.12236.

Tables

Table 4.11.1. Candidate model (m164_GSI) settings.

Age Classes	1-6+
Fleet Selectivity	Logistic
Fleet Selectivity Random Effects	iid
Fleet Age Composition	Logistic normal-ar1-miss0
Fleet CV	0.05
Survey Selectivities	NEFSC Spring & Winter: Logistic NEFSC Fall: Age-specific

Survey Selectivity Random Effects	None
Survey Catchability Random Effects	None
Survey Catchability Environmental Effects	None
Survey Age Compositions	Logistic normal-ar1-miss0
Recruitment Assumptions	Decoupled from ages 2-6+ GSI Informed (controlling effect)
Gulf Stream Index	Modeled via ar1 process
Natural Mortality Assumptions	M = 0.5 across all ages and years
Numbers-at-Age Random Effects	2dar1

Table 6.1.1. Biological reference points estimated as described in section 6.0 and terminal year (2022) values of SSB and F.

$F_{40\%}$	0.73
F_{2022}	0.11
$SSB_{40\%}$	126 mt
SSB_{2022}	44 mt
MSY proxy	97 mt

Table 7.1.1. Settings used for projections of the candidate model.

Random Effects	ar1 on age-1 (R) 2dar1 on ages 2-6+ iid on fleet selectivity
Natural Mortality	Constant (M = 0.5)
Weight-at-Age	Terminal 2-year average
Maturity	Constant
Recruitment	GSI informed
GSI	Mean 2012-2022

Table 7.2.1. Estimates and uncertainties (90% and 95% confidence intervals) of four years of projected Catch (mt), F, R (000s), and SSB (mt). Forecasts were done using bridge year (2023) catch equal to 2022 catch and then fishing at $F_{40\%} = 0.73$ in years 2024-2026.

Type	*Year*	*Estimation*	*Low 90*	*High 90*	*Low 95*	*High 95*
Catch	2023	5	5	5	5	5
Catch	2024	31	9	106	7	135
Catch	2025	50	8	322	5	460
Catch	2026	67	8	572	5	863
F	2023	0.13	0.05	0.34	0.04	0.42
F	2024	0.73	0.54	0.97	0.51	1.03
F	2025	0.73	0.54	0.97	0.51	1.03
F	2026	0.73	0.54	0.97	0.51	1.03
SSB	2023	35	13	96	11	117
SSB	2024	35	9	129	7	165
SSB	2025	56	9	362	6	517
SSB	2026	79	9	702	6	1067
R	2023	659	168	2581	129	3352
R	2024	1428	189	10799	128	15911
R	2025	1429	189	10818	128	15941
R	2026	1429	189	10819	128	15944

Figures

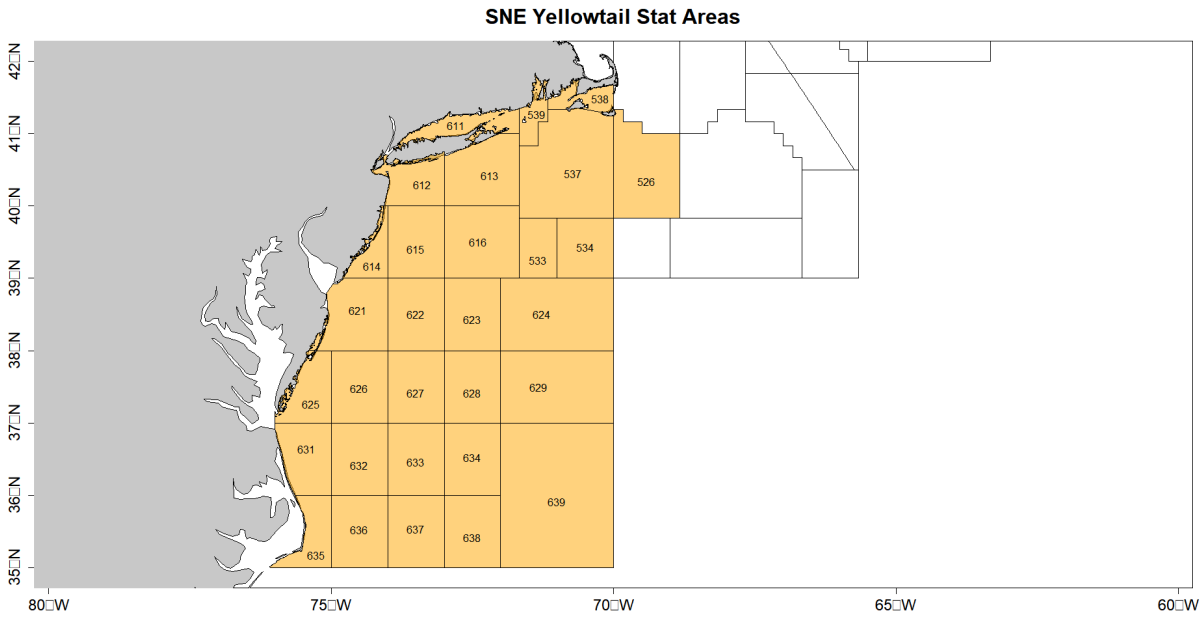

Figure 2.0.1. Statistical areas that comprise the southern New England / Mid-Atlantic yellowtail flounder stock.

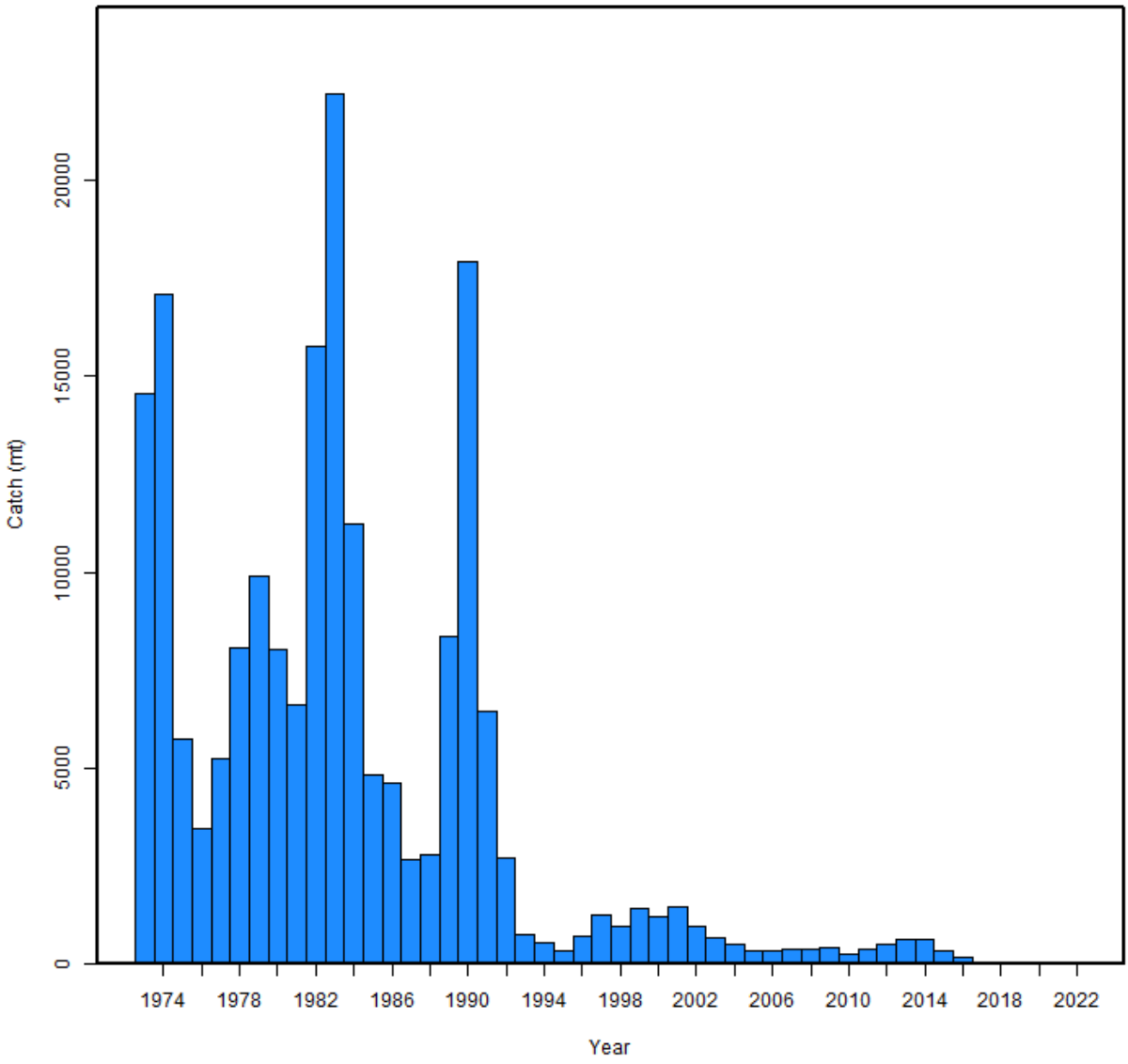

Figure 3.1.1 Total removals of yellowtail flounder from the southern New England / Mid-Atlantic stock region. This represents a combined index of landings and discards. Note that catches in 2018-2022 occurred, but are at relatively low values compared to the historical series.

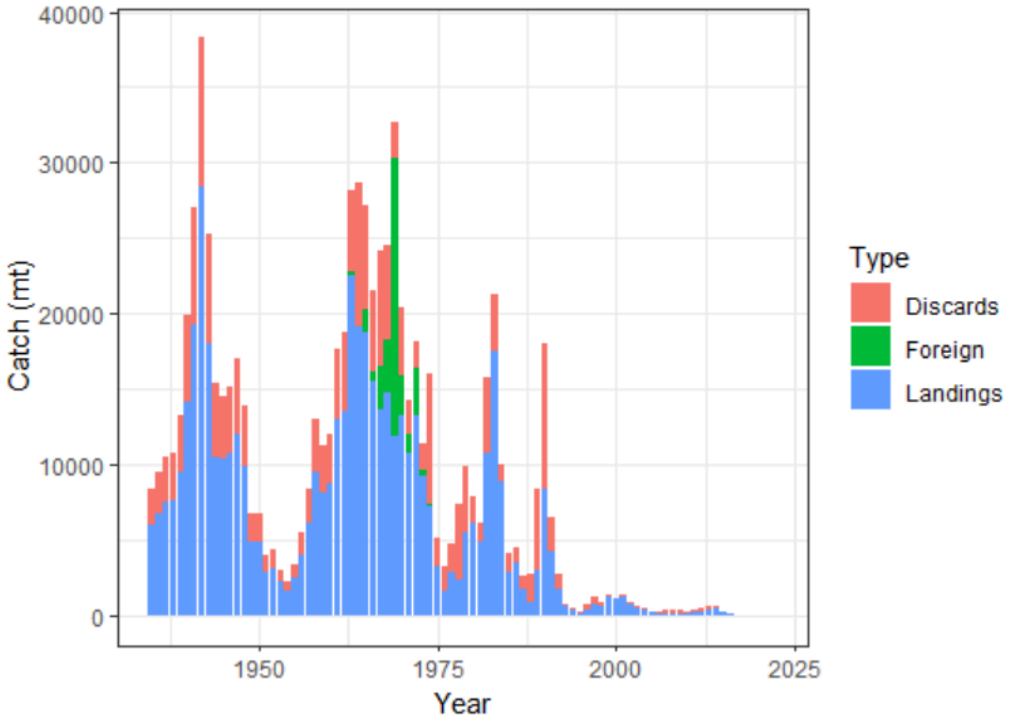

Figure 3.1.2. Total removals of yellowtail flounder from the southern New England / Mid-Atlantic stock region broken up into discards, landings, and foreign landings.

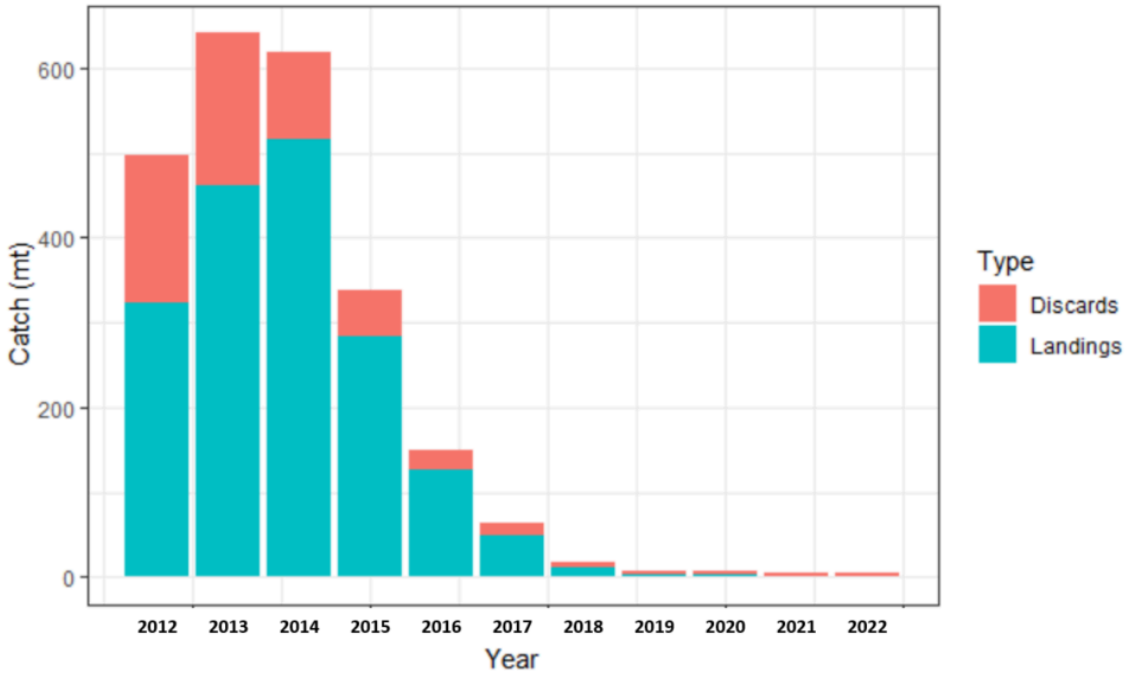

Figure 3.1.3. Total removals of yellowtail flounder from the southern New England / Mid-Atlantic stock region 2012-2022 broken up into discards and landings.

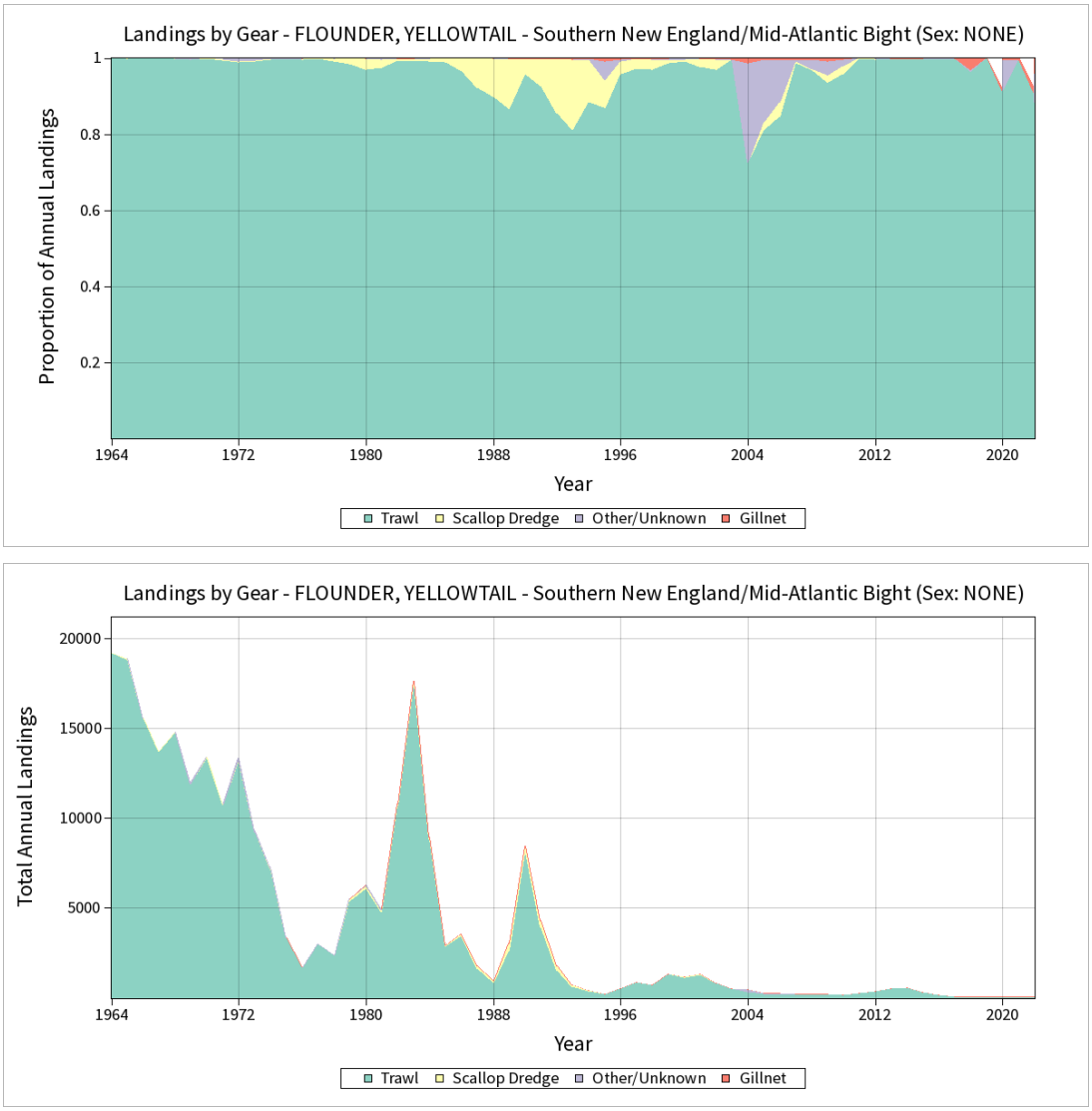

Figure 3.1.4. Proportion of total landings of yellowtail flounder from the southern New England / Mid-Atlantic stock region broken up by gear type (top) and total removals by gear type (bottom).

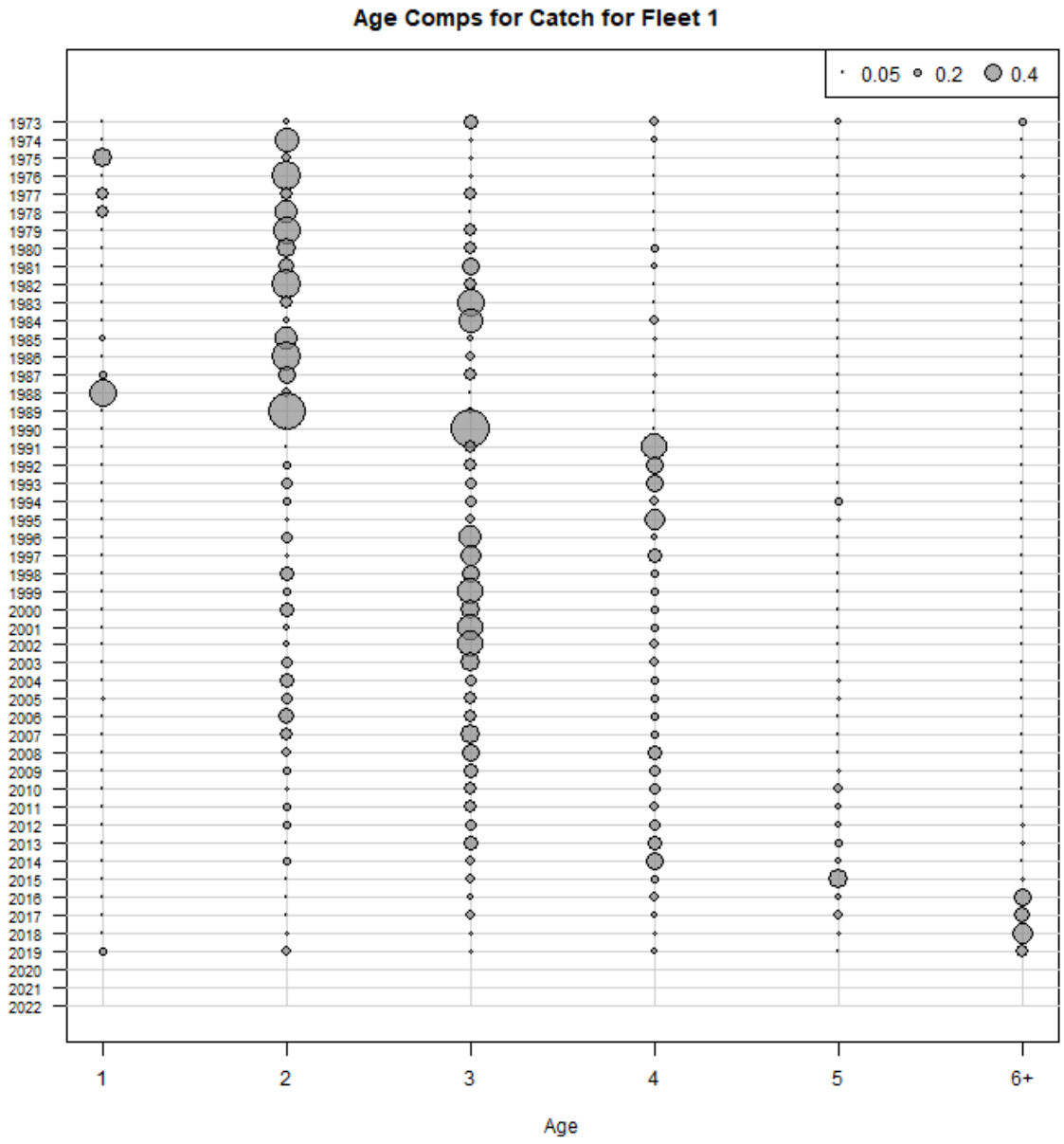

Figure 3.1.5. Age composition of the combined fishery catch.

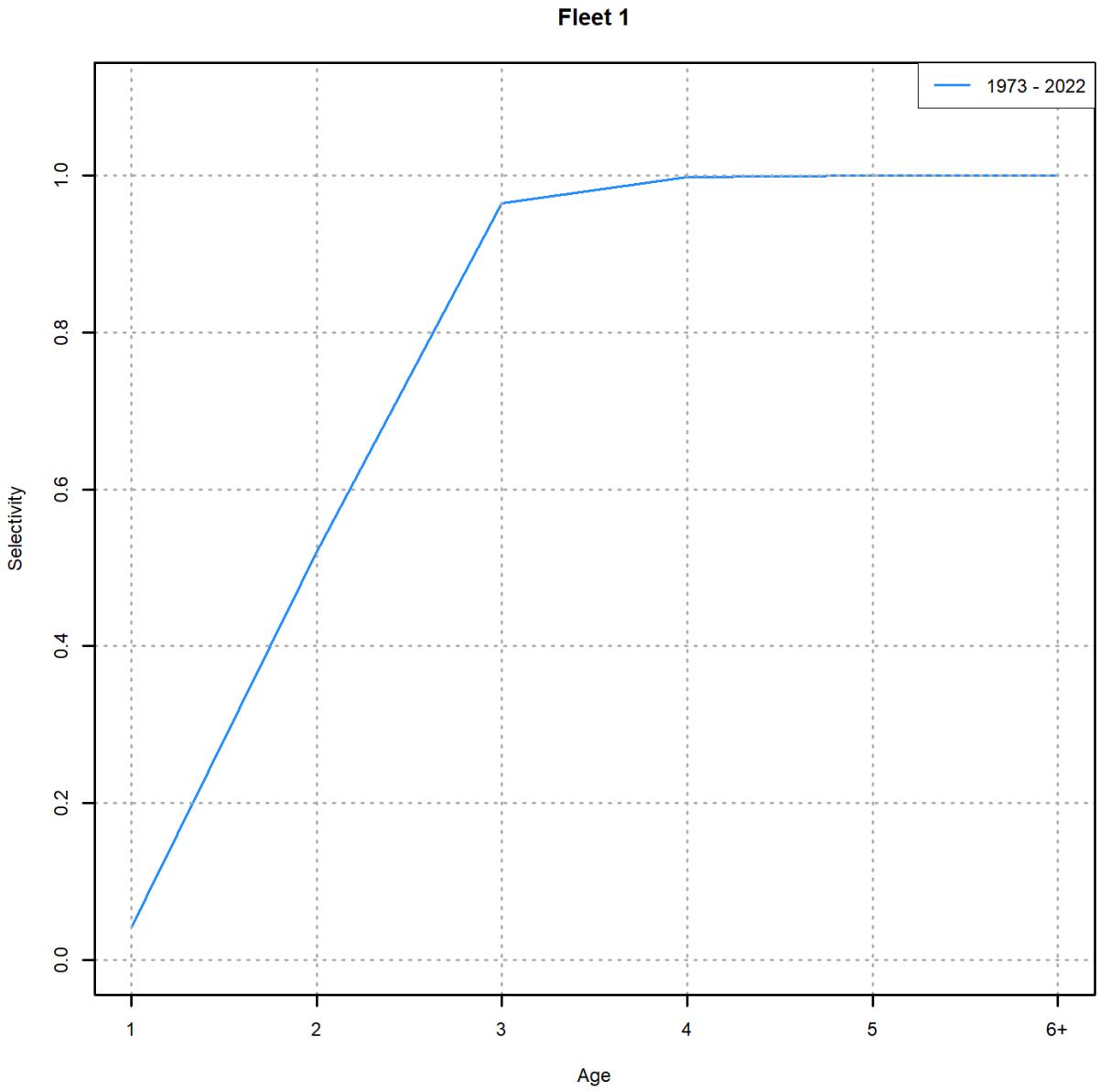

Figure 4.2.1. Combined fleet selectivity in the candidate model averaged across the series 1973-2022. Fleet selectivity in the candidate model varies 1973-2022 due to the iid random effects. Figure 5.2.4 shows these changes over the series.

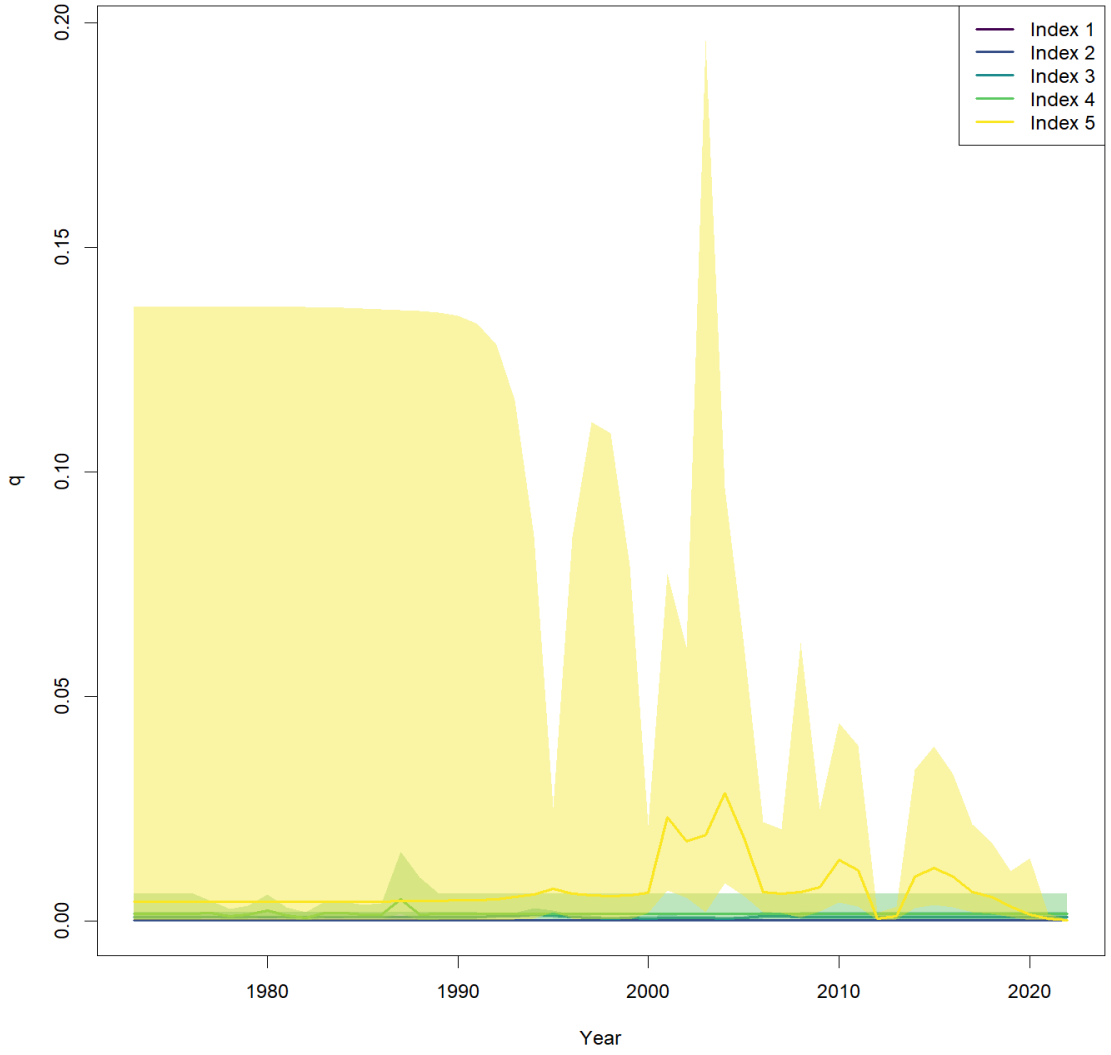

Figure 4.7.1. Catchability over time for each of the survey indices. These results are NOT from the candidate model, as the candidate model does not include the larval indices. These catchabilities are only included to show how unrealistic the catchability estimates for the larval indices are if they were to be included. Indices 1-3 are the NEFSC spring, fall, and winter indices, respectively, and indices 4-5 are the MARMAP and ECOMON larval indices, respectively.

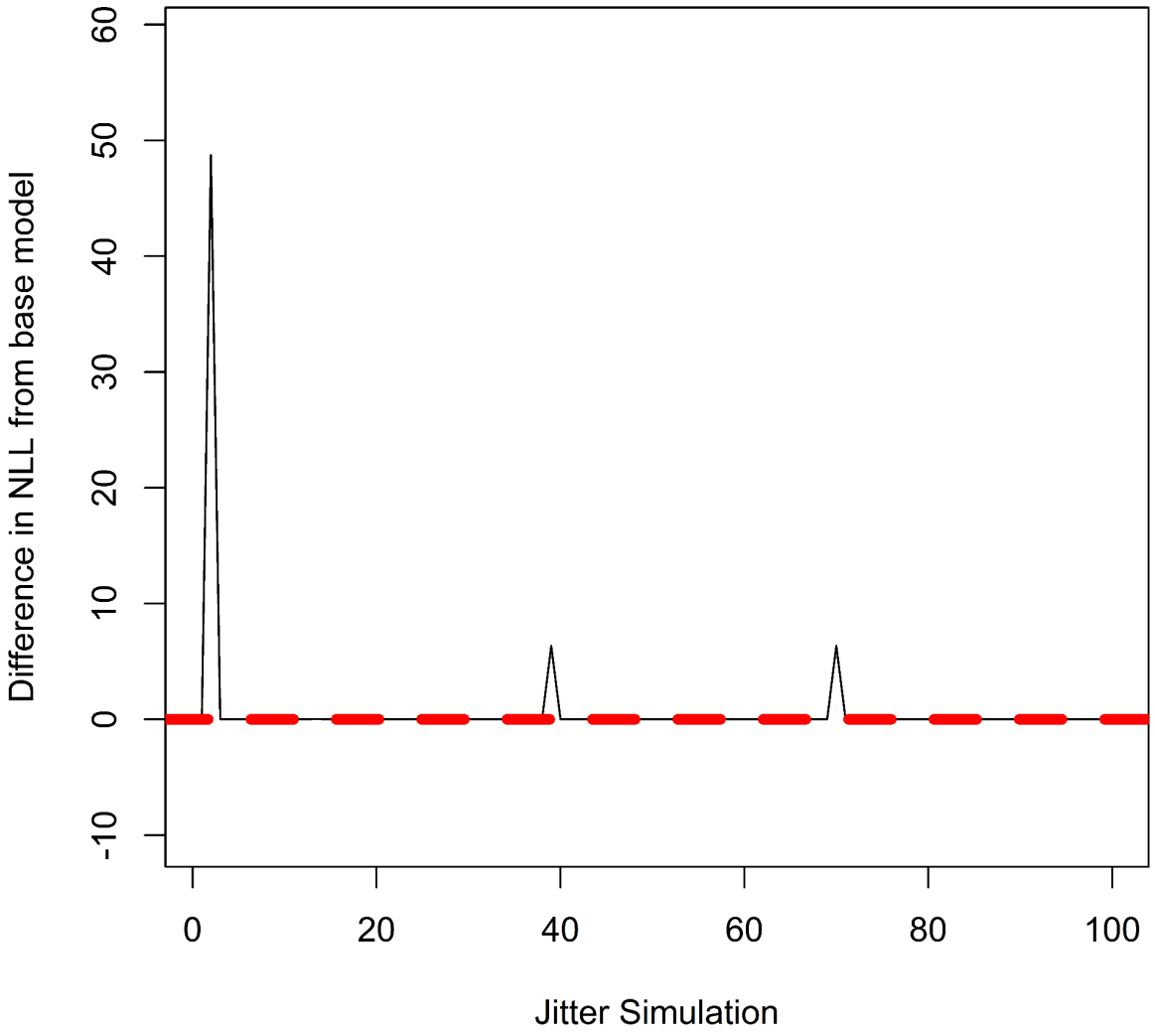

Figure 5.1.1. Jitter analysis results of the candidate model. Convergence rate was 93% and all but three runs (of 100) converged to the global minimum, indicating high stability.

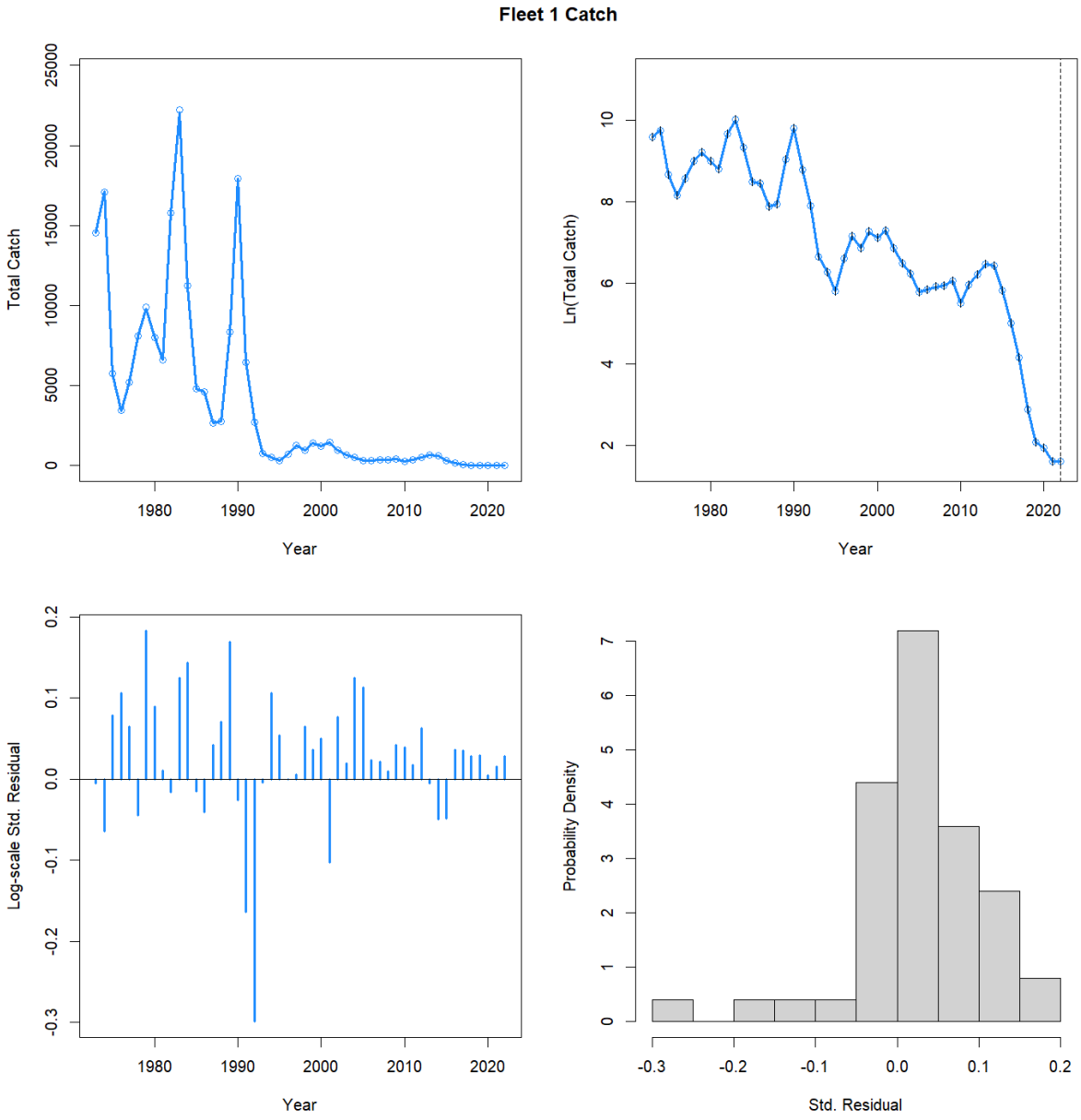

Figure 5.1.2. Candidate model fit to the aggregate fleet data.

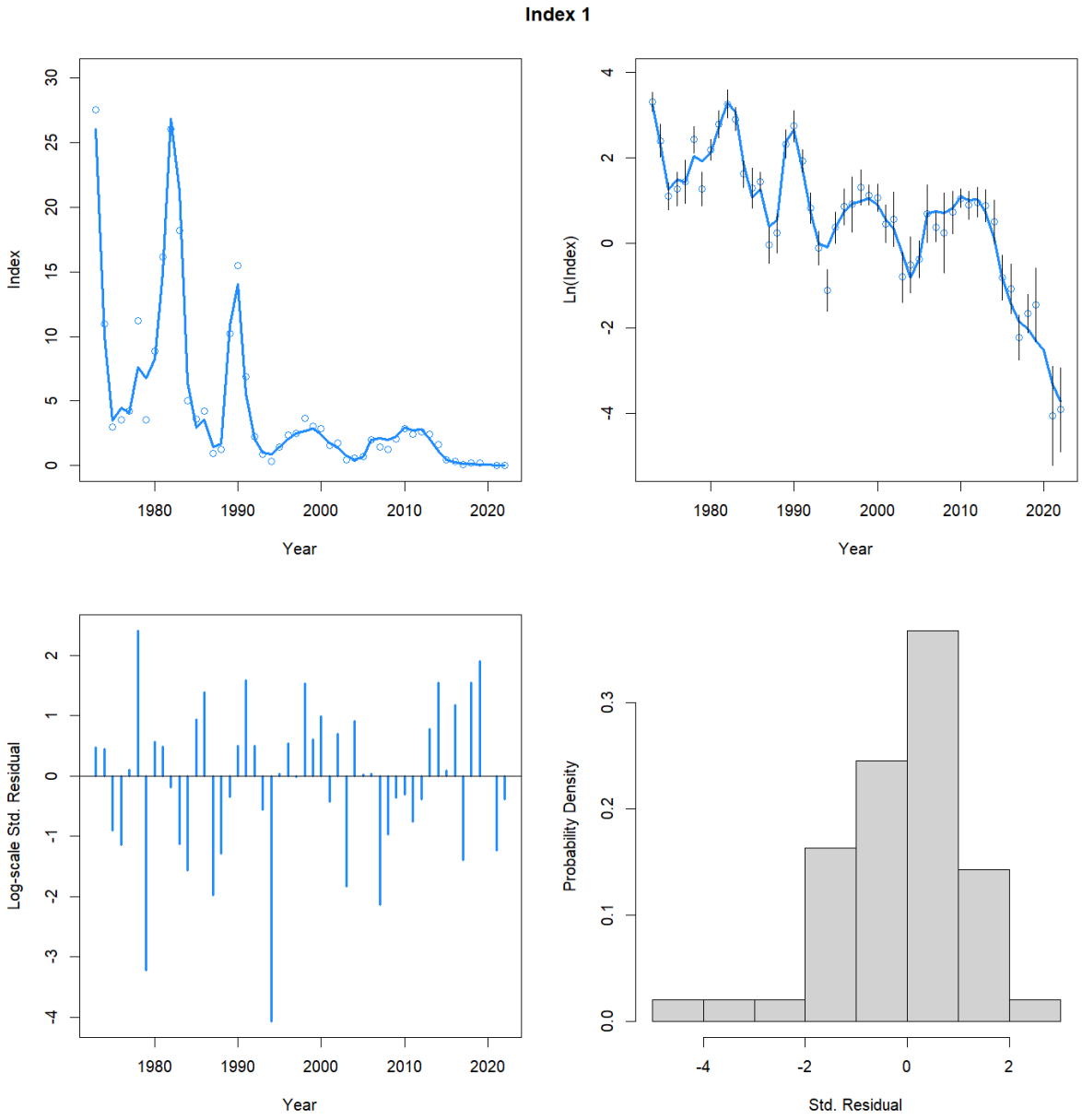

Figure 5.1.3. Candidate model fit to the NEFSC spring bottom trawl survey index.

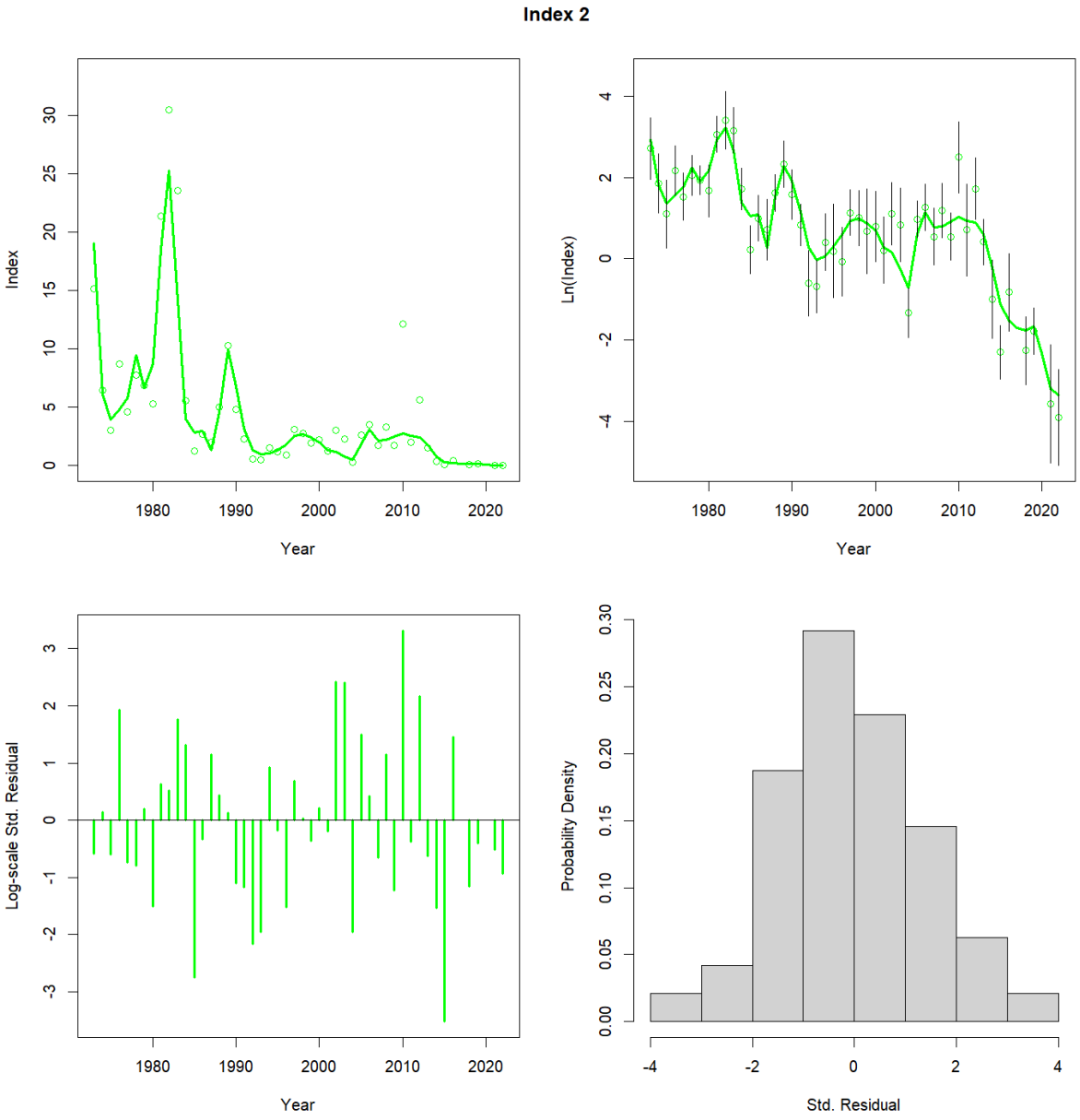

Figure 5.1.4. Candidate model fit to the NEFSC fall bottom trawl survey index.

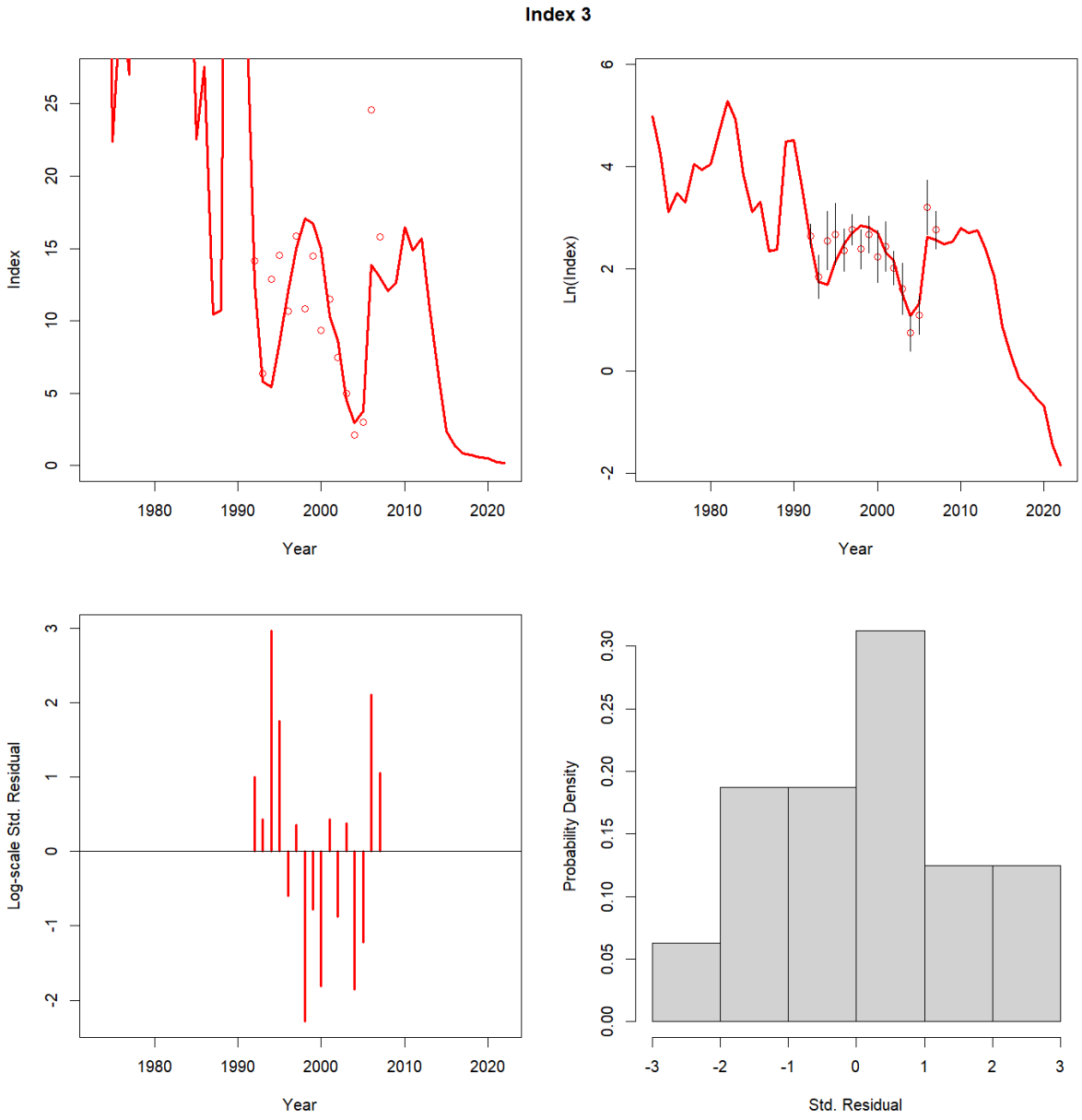

Figure 5.1.5. Candidate model fit to the NEFSC winter bottom trawl survey index.

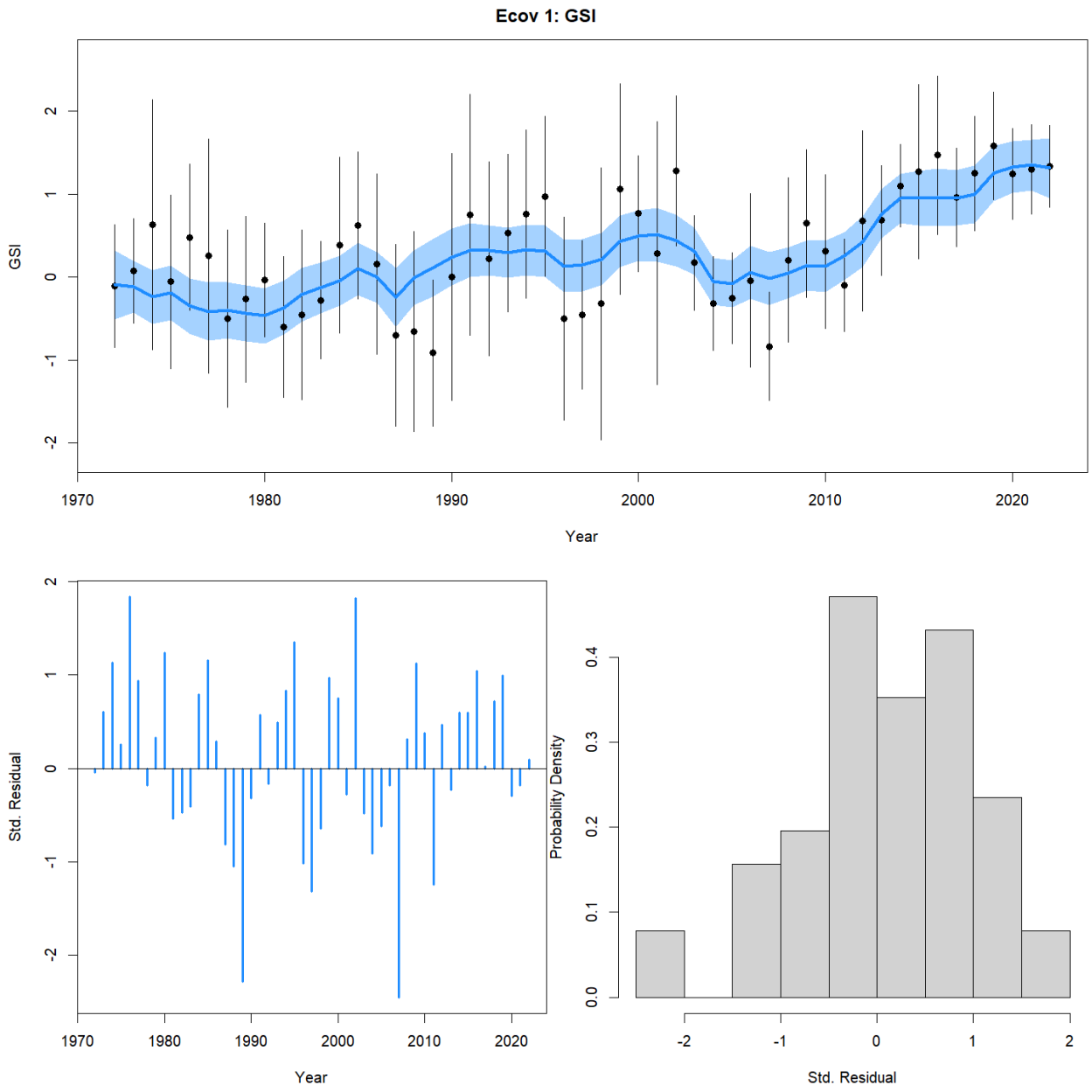

Figure 5.1.6. Candidate model fit to the GSI.

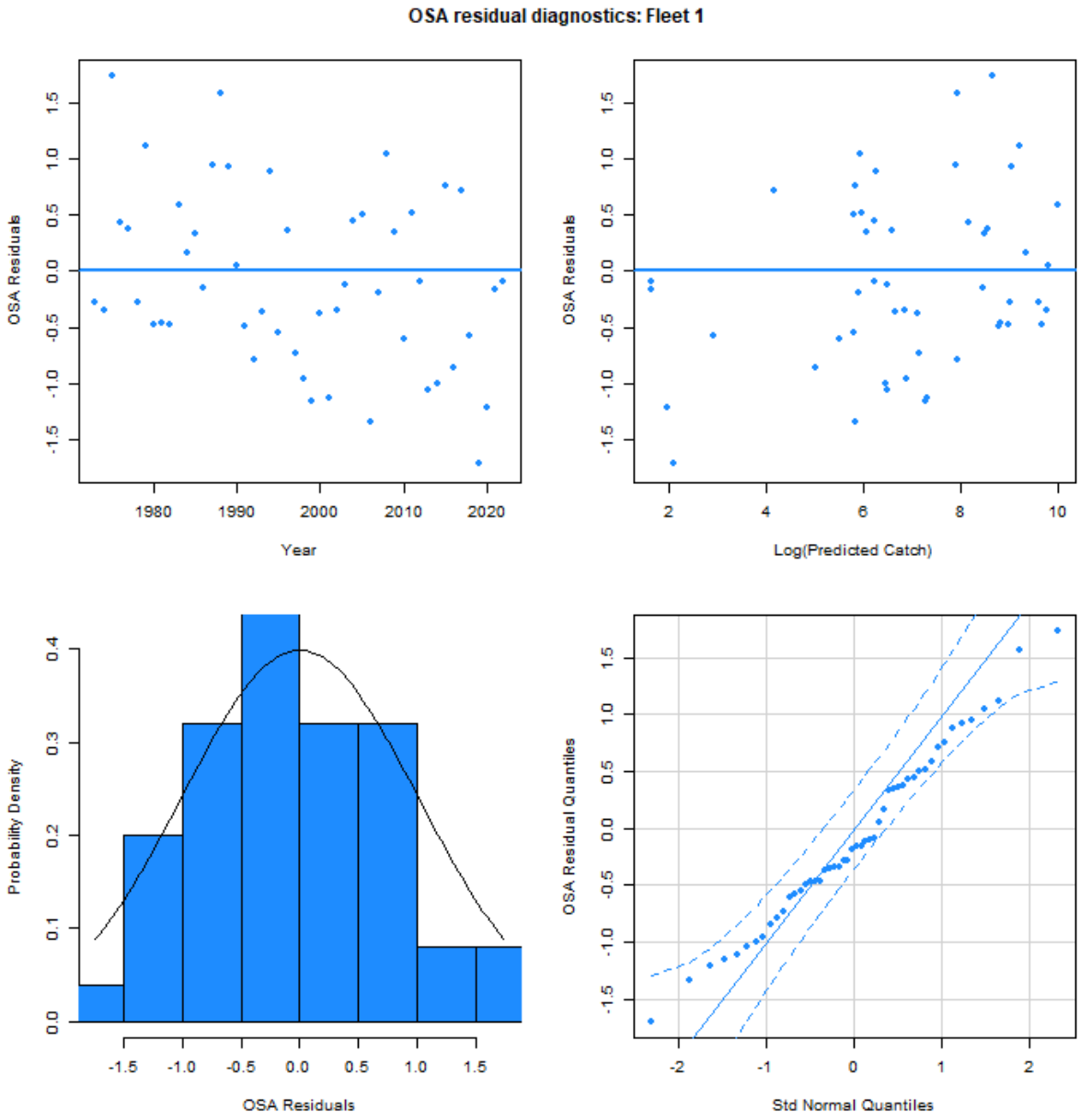

Figure 5.1.7. OSA residual diagnostics for the aggregate fleet.

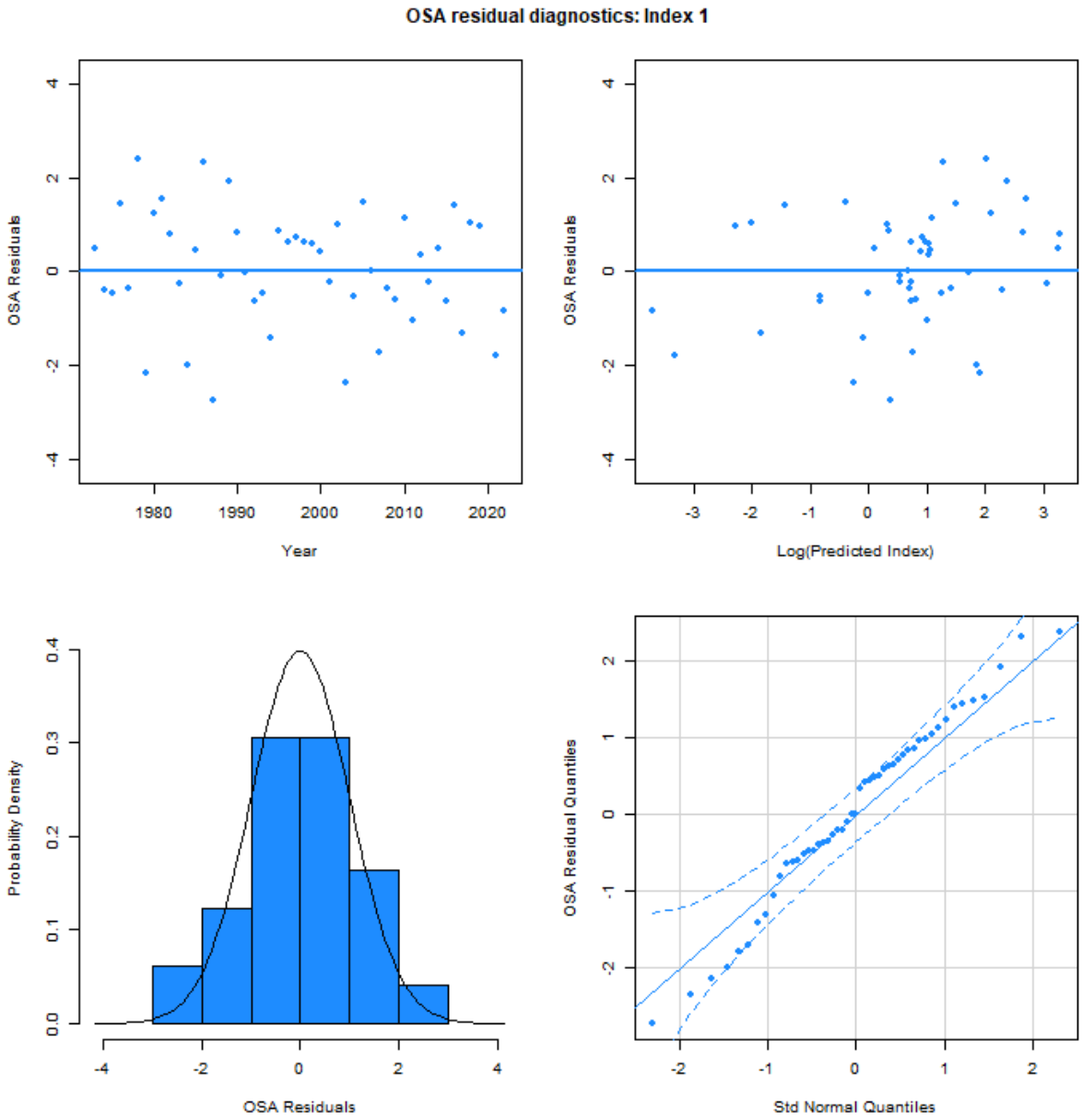

Figure 5.1.8. OSA residual diagnostics for the NEFSC spring bottom trawl index.

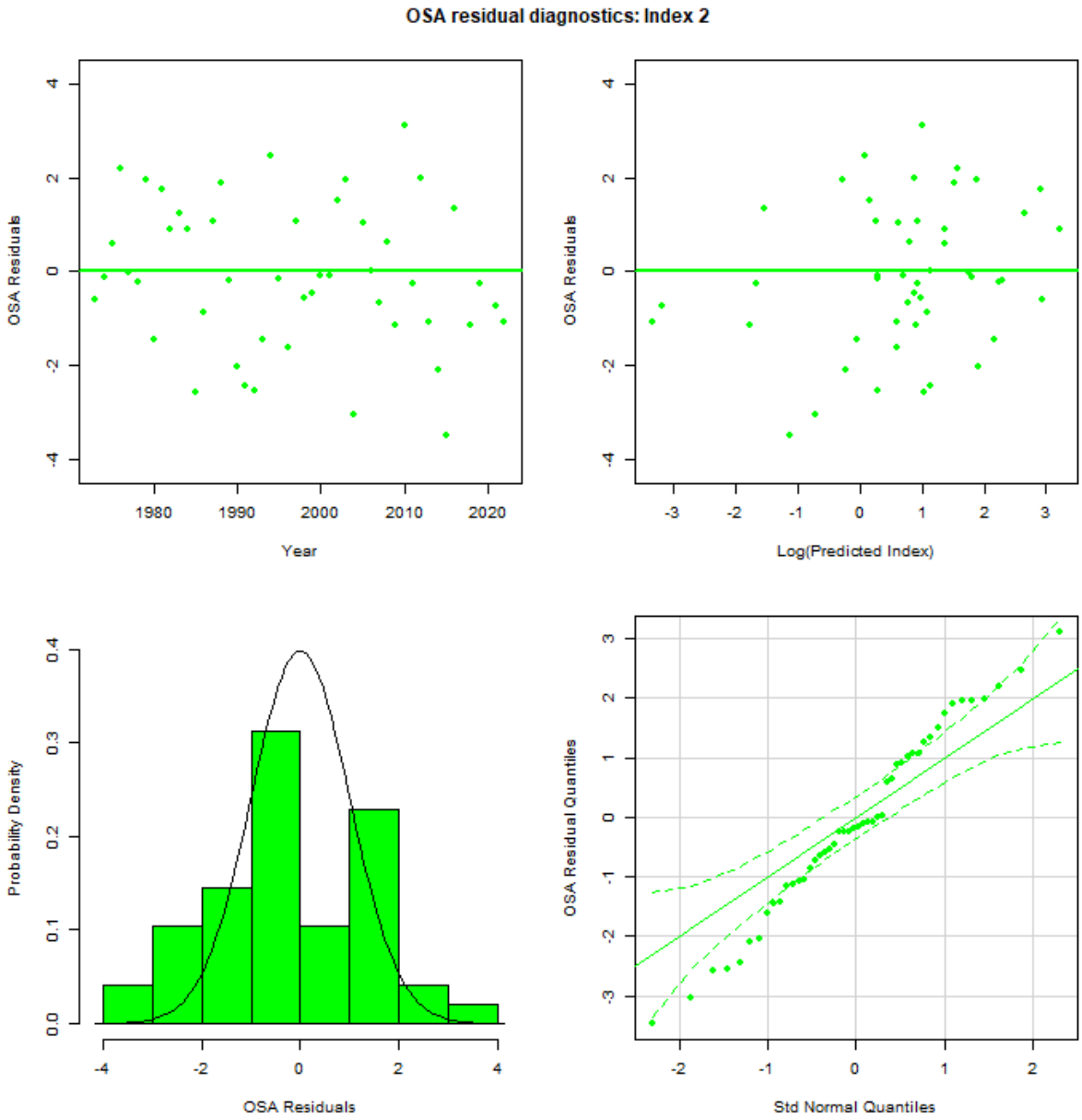

Figure 5.1.9. OSA residual diagnostics for the NEFSC fall bottom trawl index.

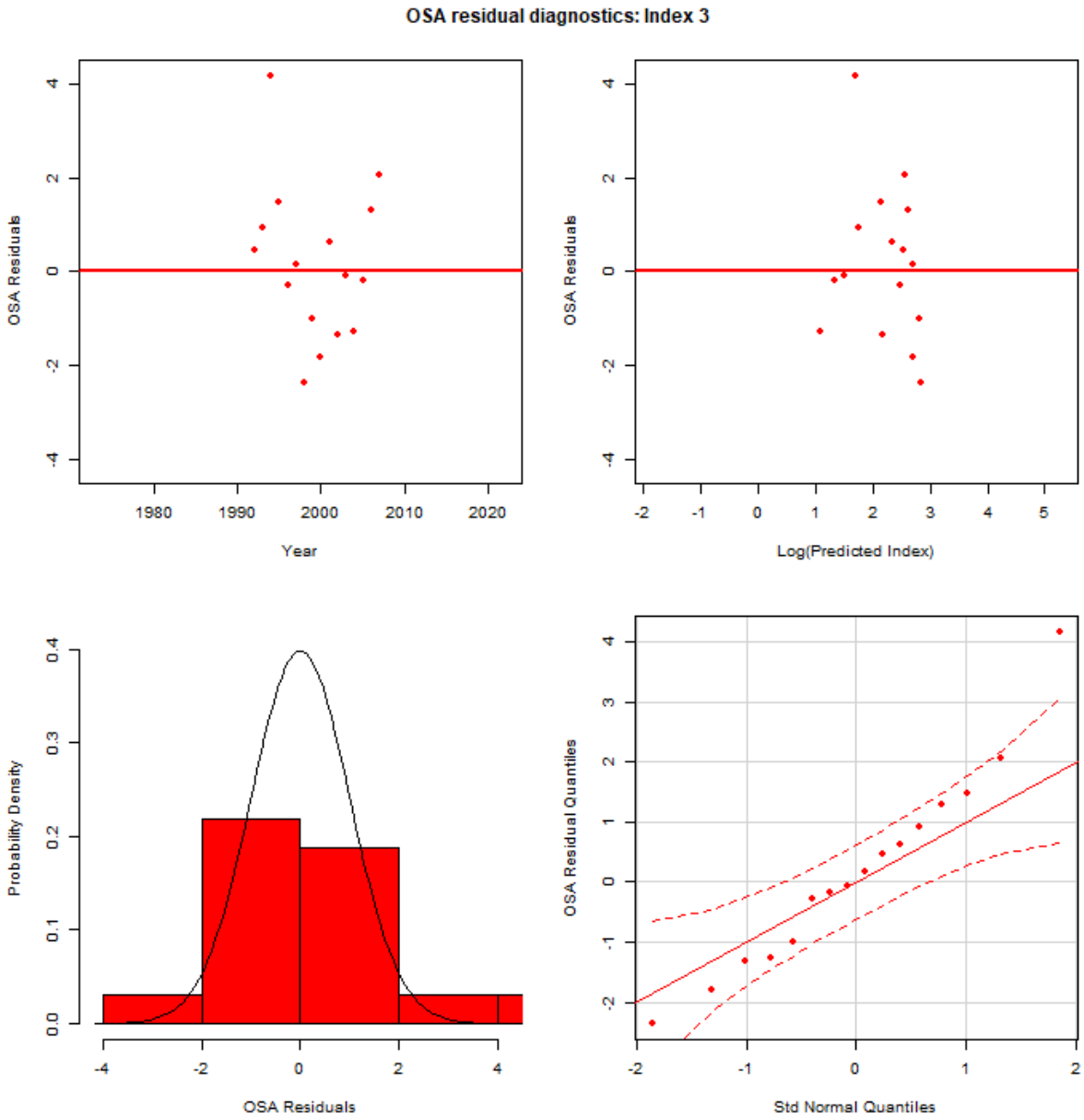

Figure 5.1.10. OSA residual diagnostics for the NEFSC winter bottom trawl index.

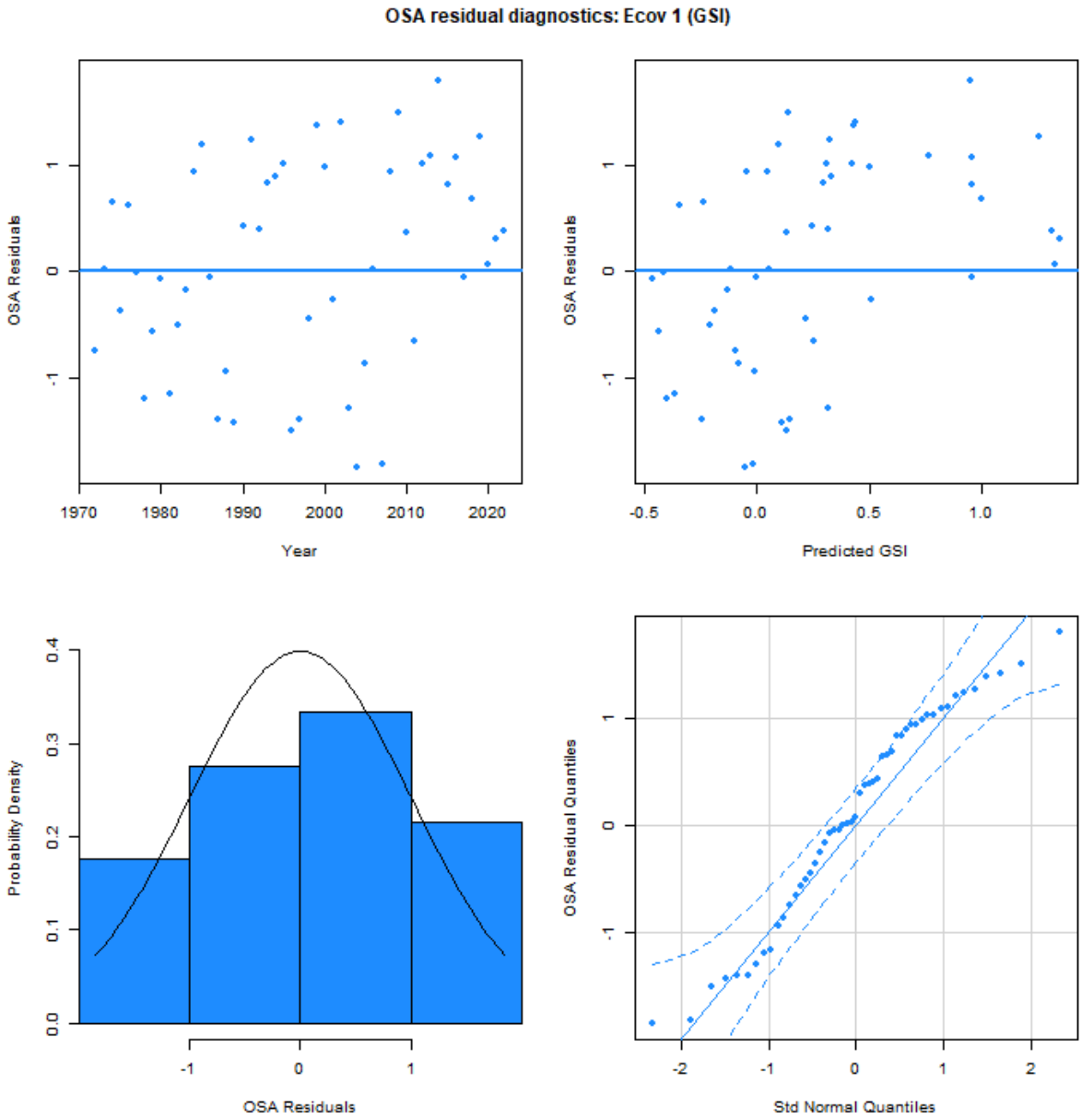

Figure 5.1.11. OSA residual diagnostics for the GSI.

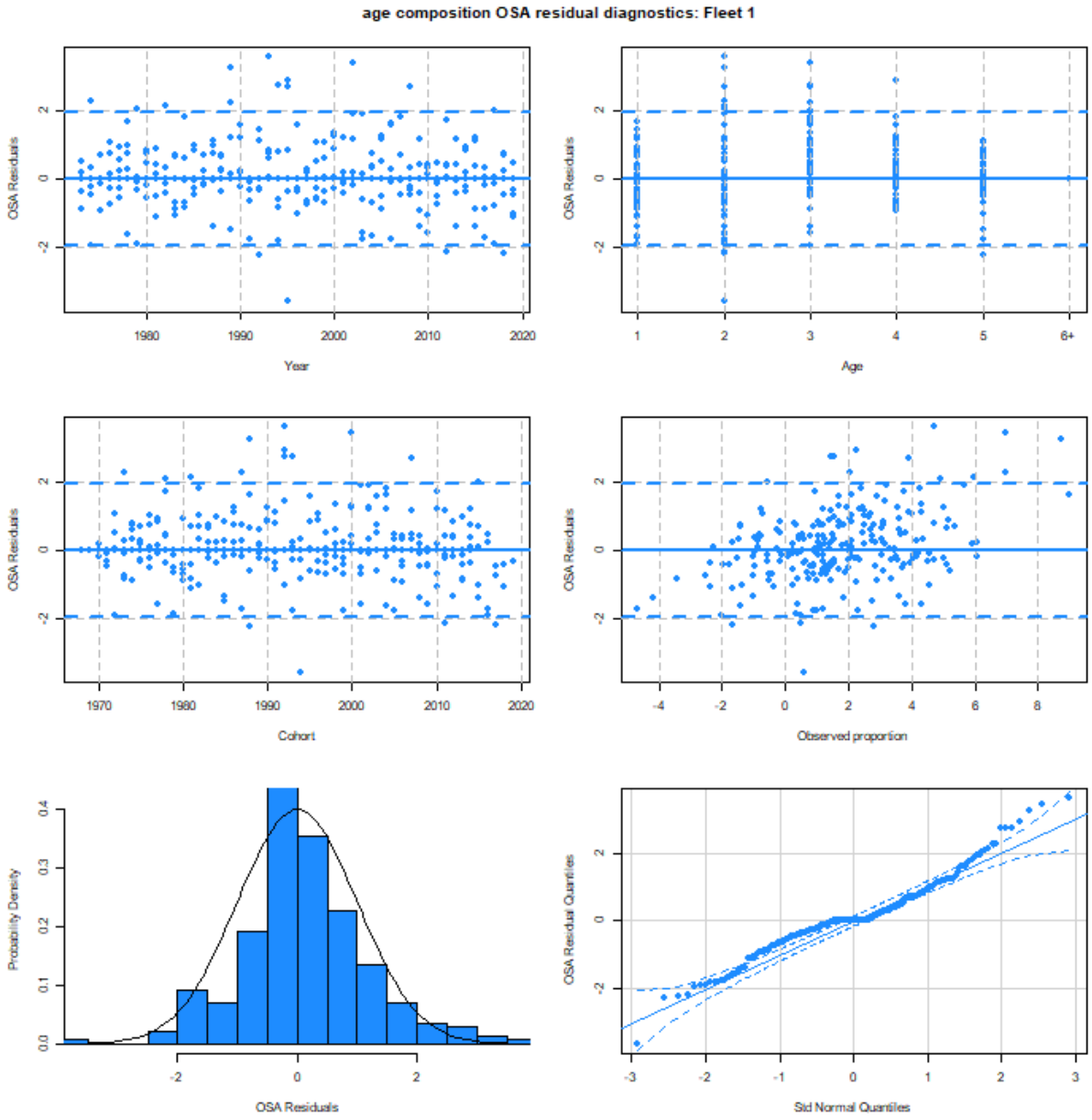

Figure 5.1.12. OSA residual diagnostics for the aggregate fleet age composition.

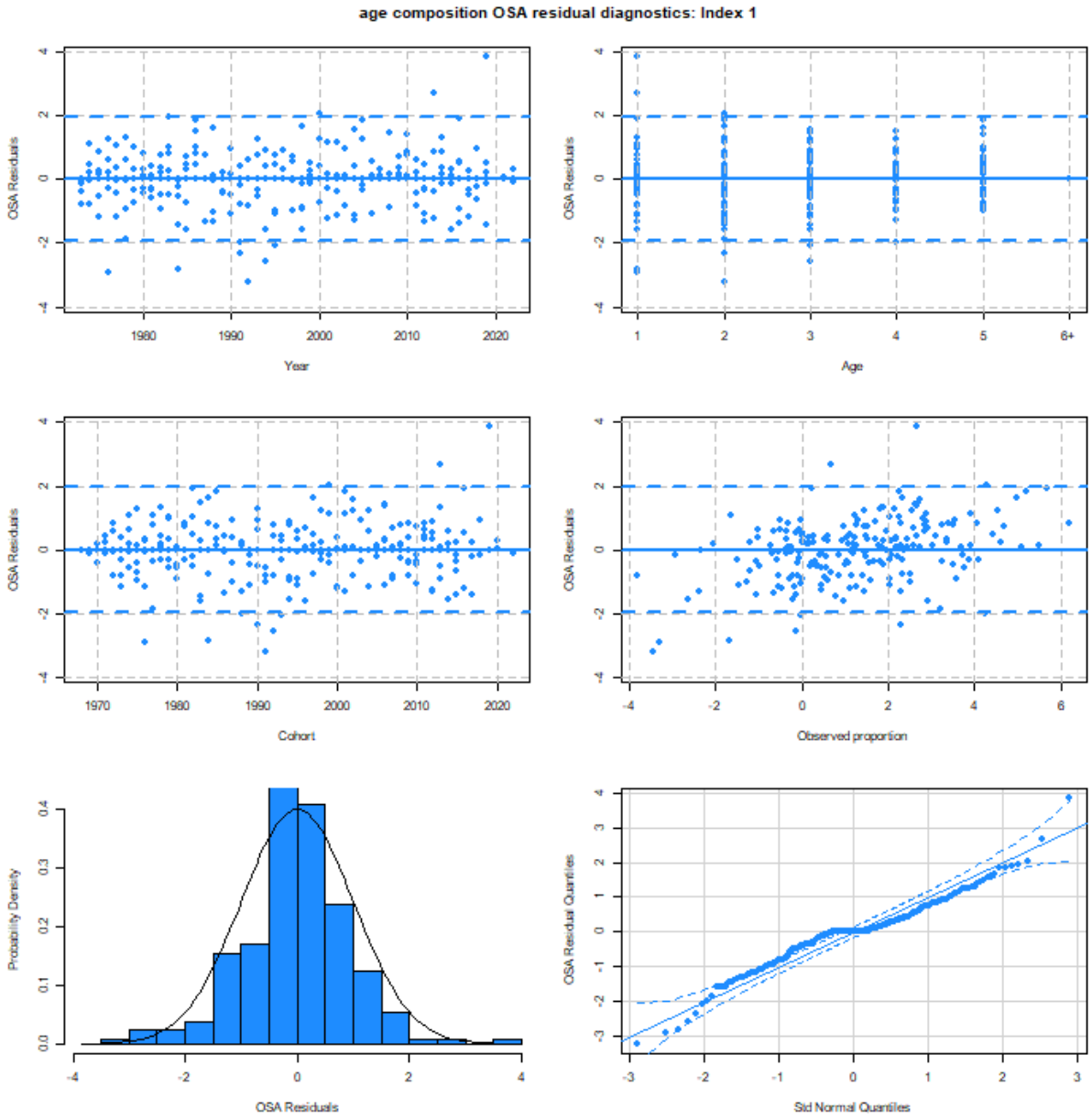

Figure 5.1.13. OSA residual diagnostics for the NEFSC spring bottom trawl index age composition.

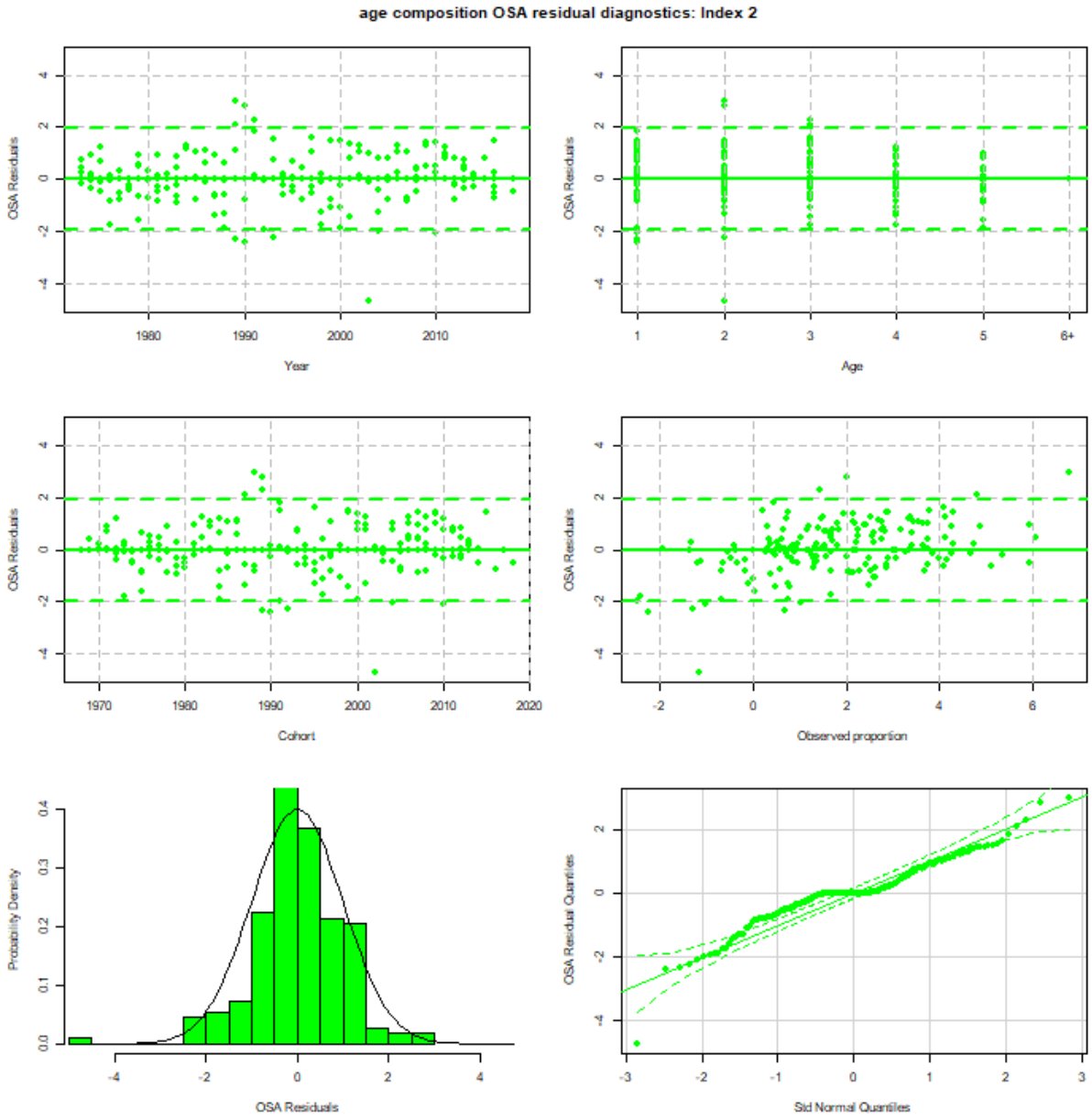

Figure 5.1.14. OSA residual diagnostics for the NEFSC fall bottom trawl index age composition.

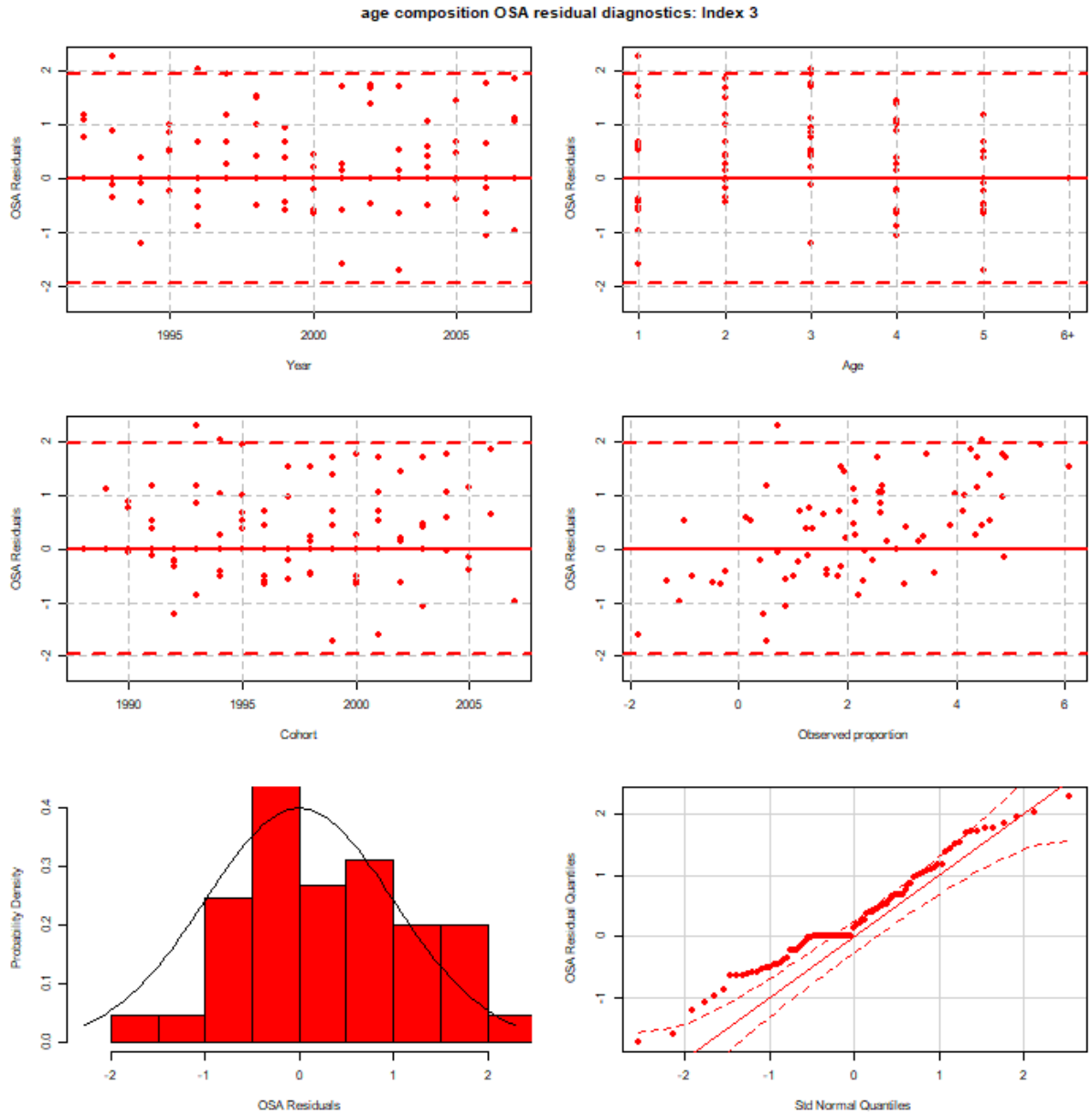

Figure 5.1.15. Bubble plot of OSA residual diagnostics for the NEFSC winter bottom trawl index age composition.

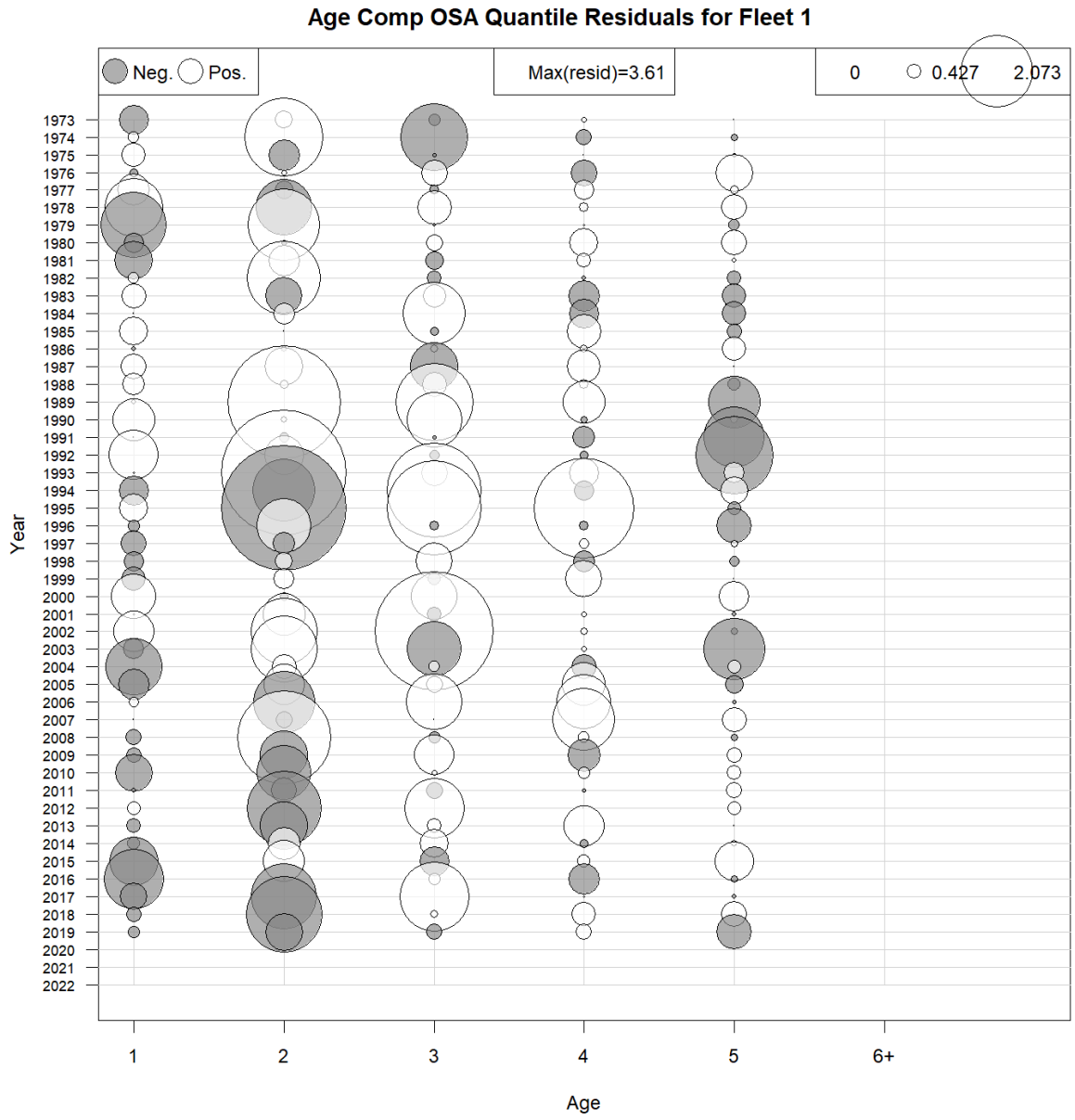

Figure 5.1.16. Bubble plot of OSA residual diagnostics for the aggregate fleet age composition.

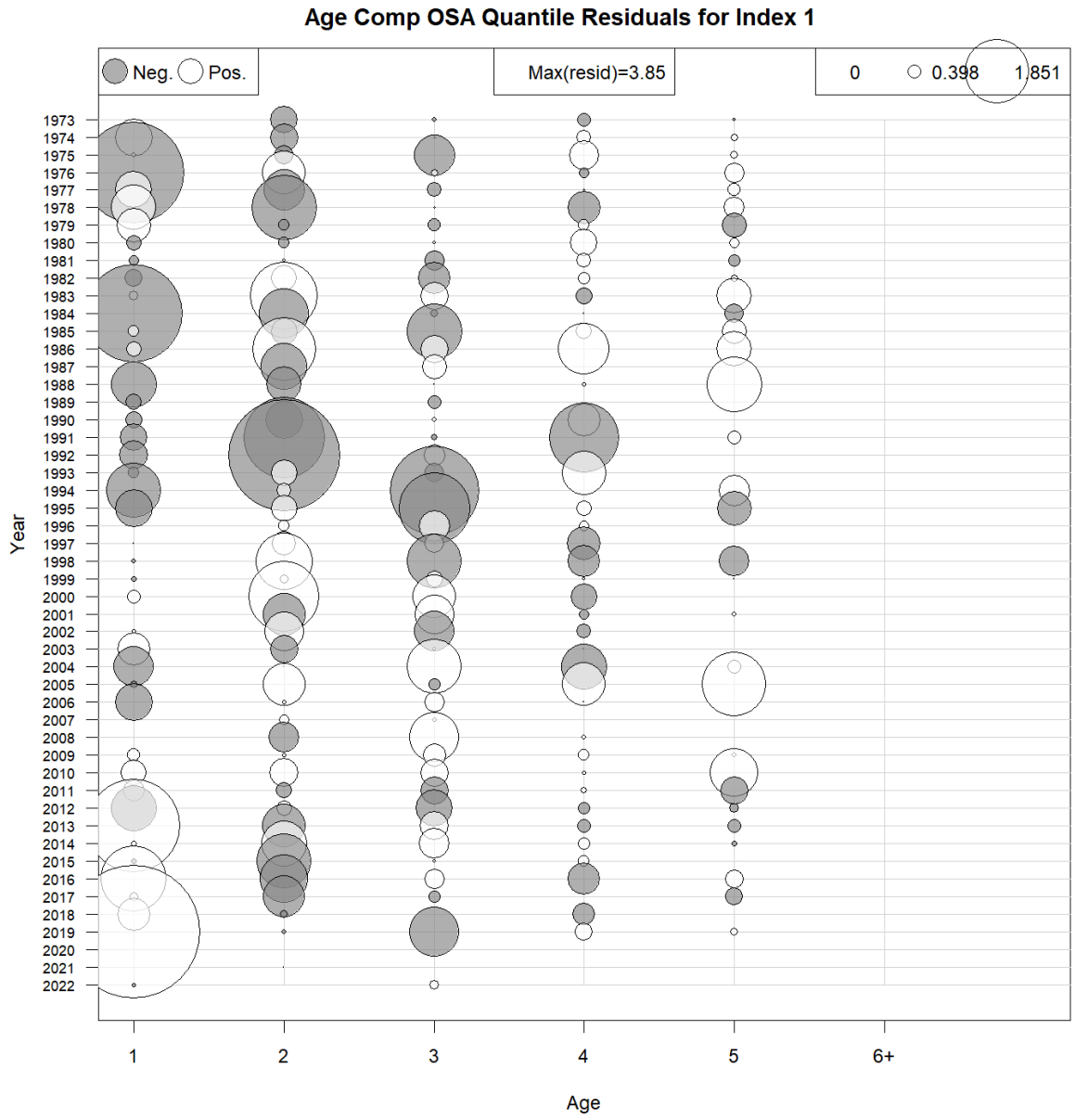

Figure 5.1.17. Bubble plot of OSA residual diagnostics for the NEFSC spring bottom trawl index age composition.

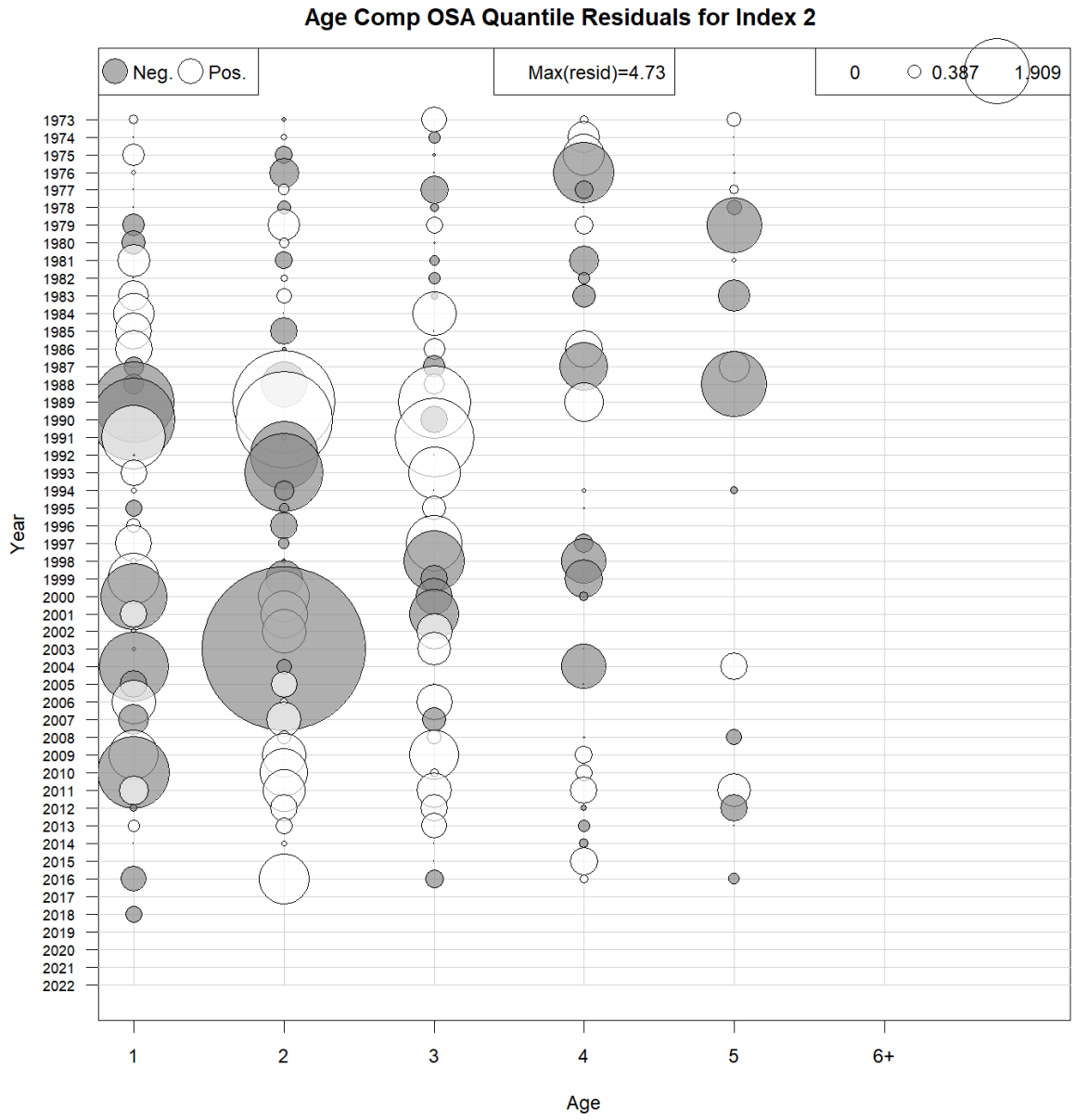

Figure 5.1.18. Bubble plot of OSA residual diagnostics for the NEFSC fall bottom trawl index age composition.

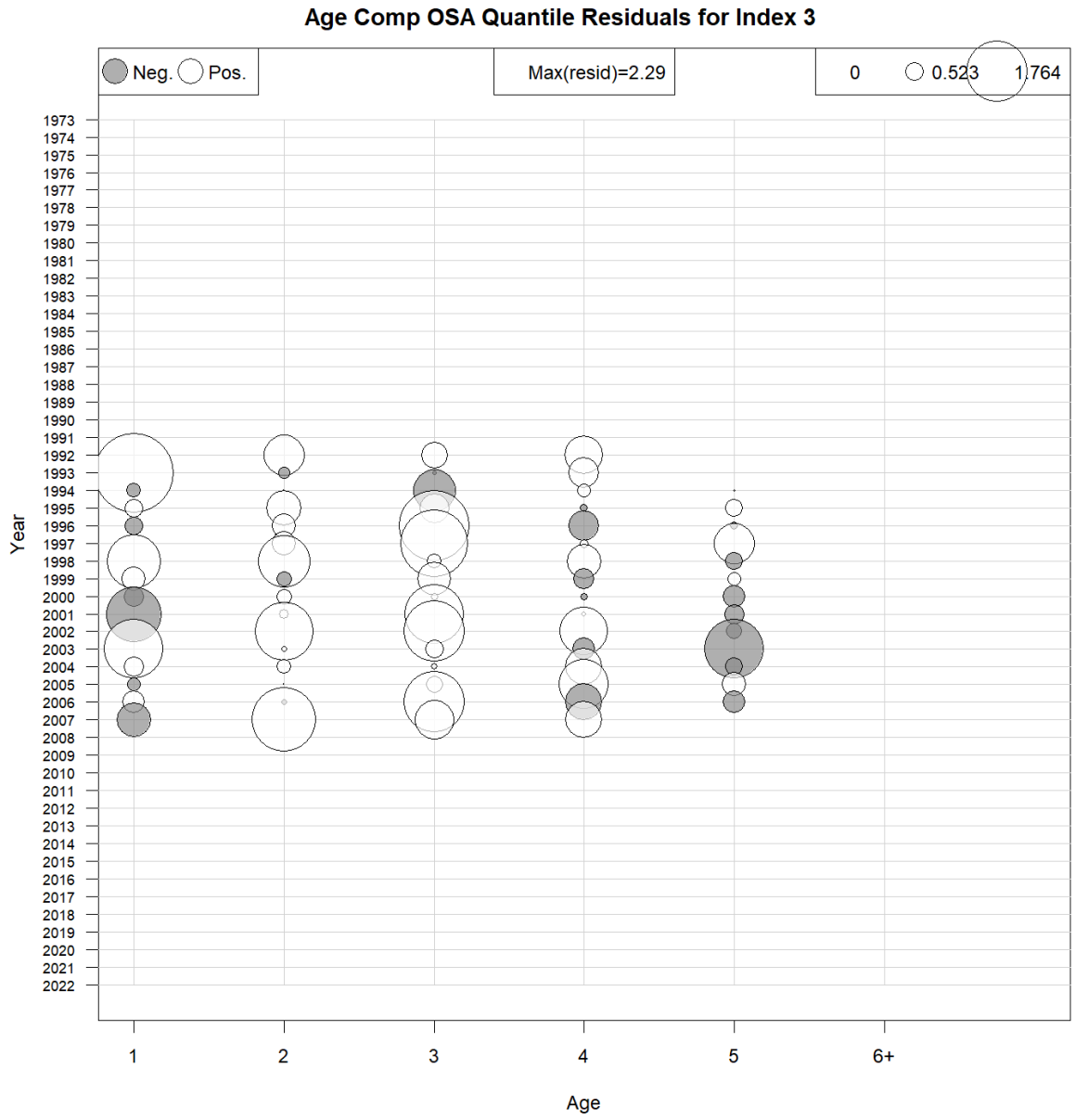

Figure 5.1.19. Bubble plot of OSA residual diagnostics for the NEFSC winter bottom trawl index age composition.

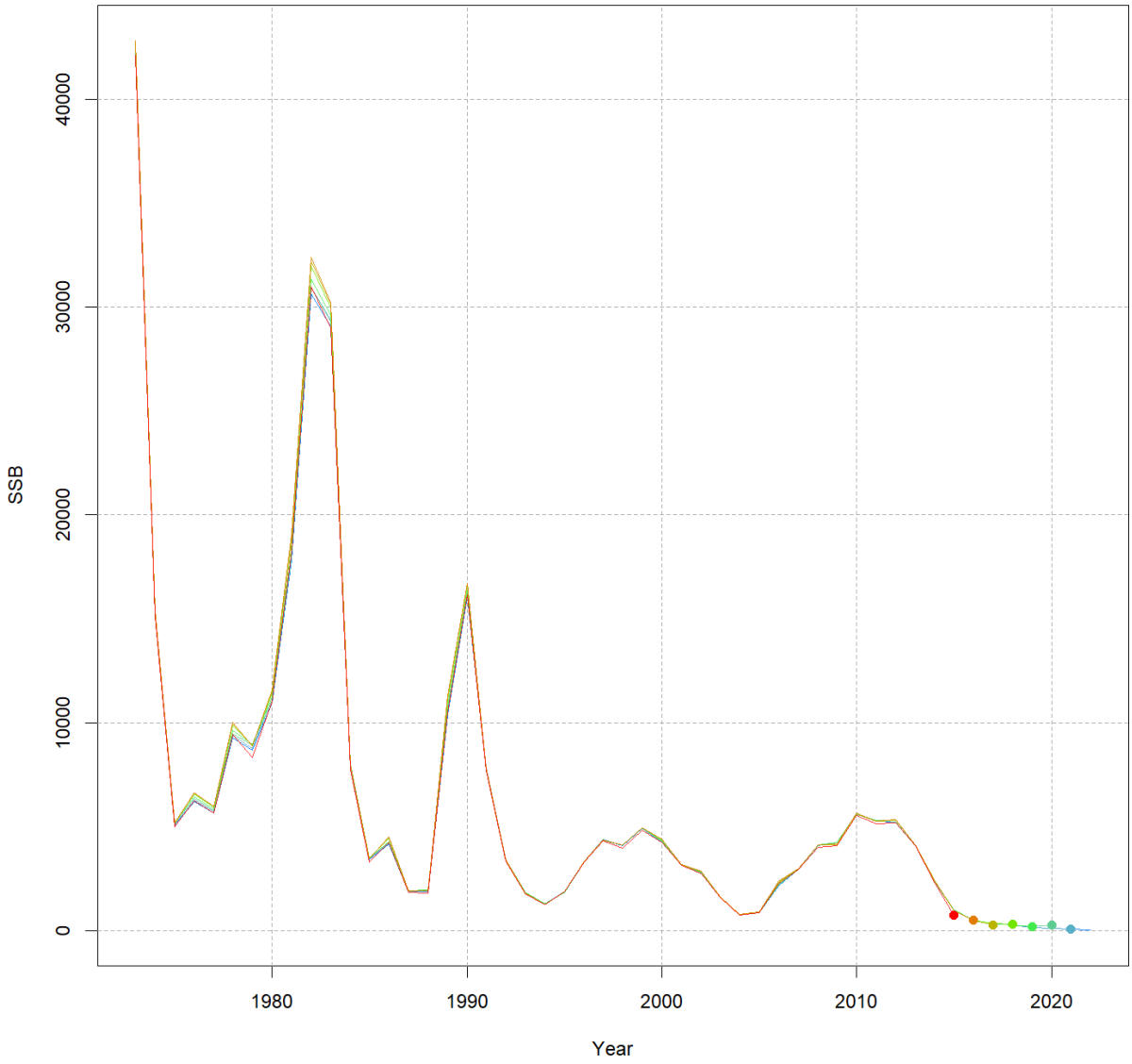

Figure 5.1.20. Retrospective pattern of SSB for the candidate model.

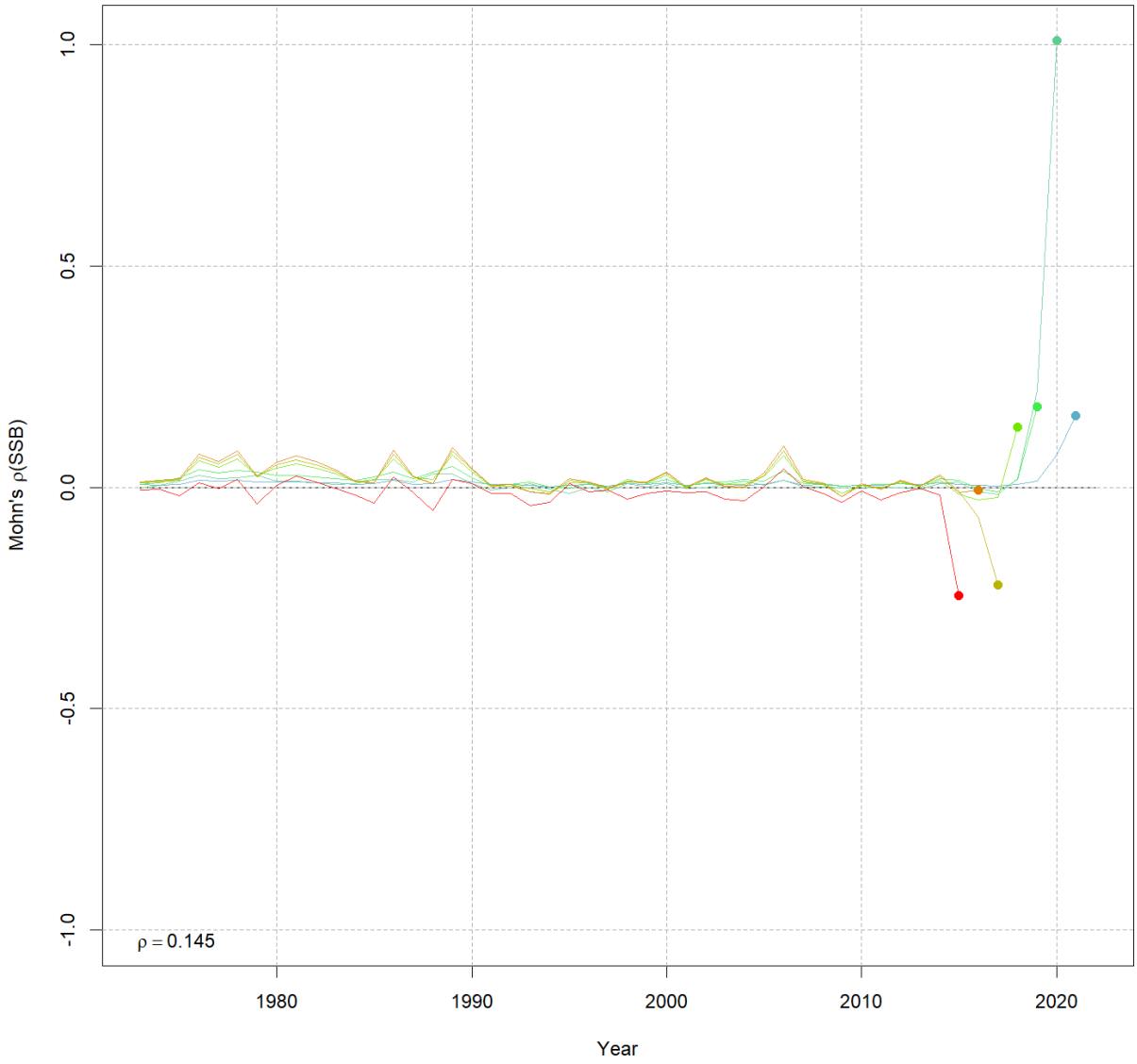

Figure 5.1.21. Relative retrospective pattern of SSB for the candidate model.

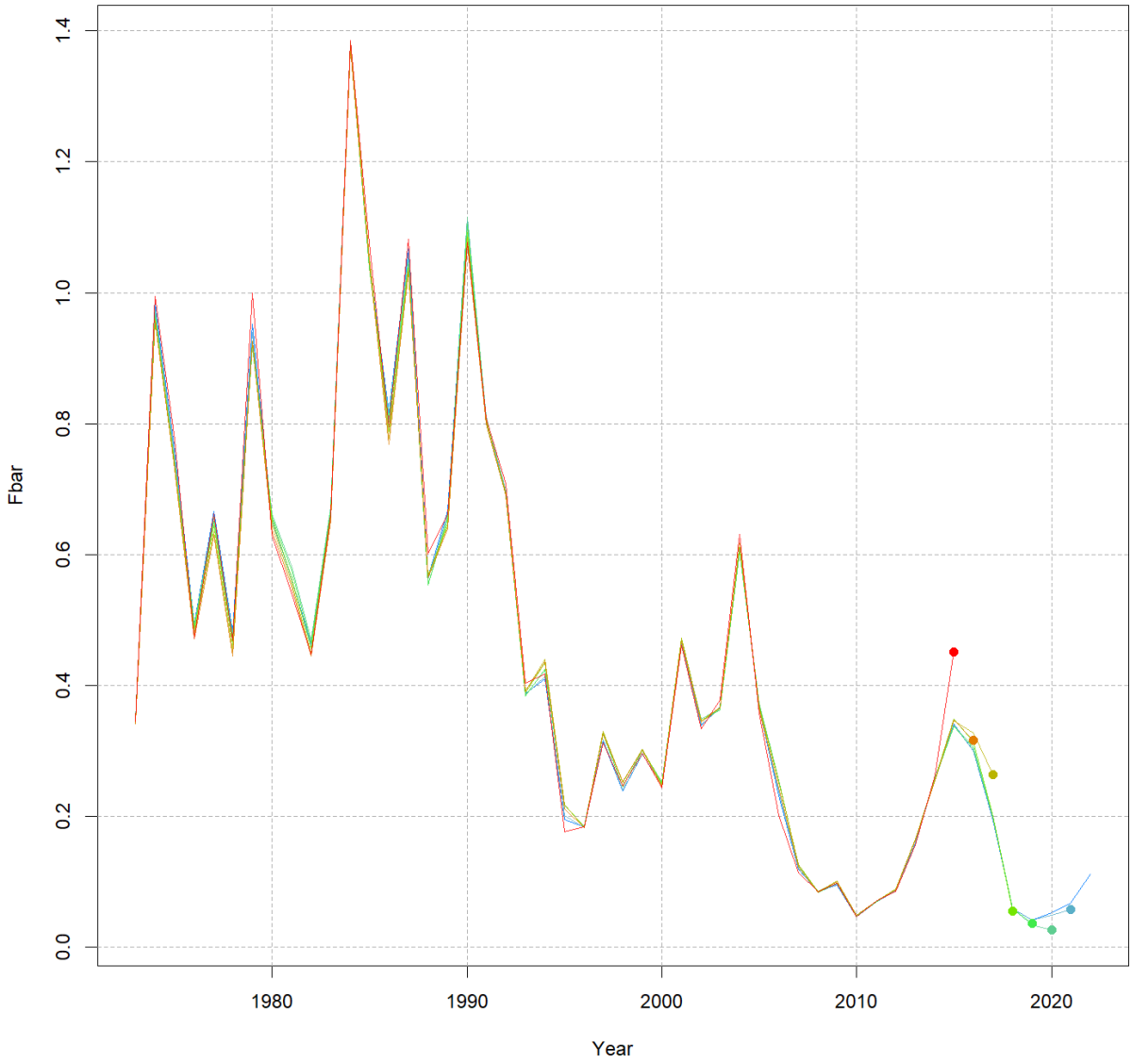

Figure 5.1.22. Retrospective pattern of F for the candidate model.

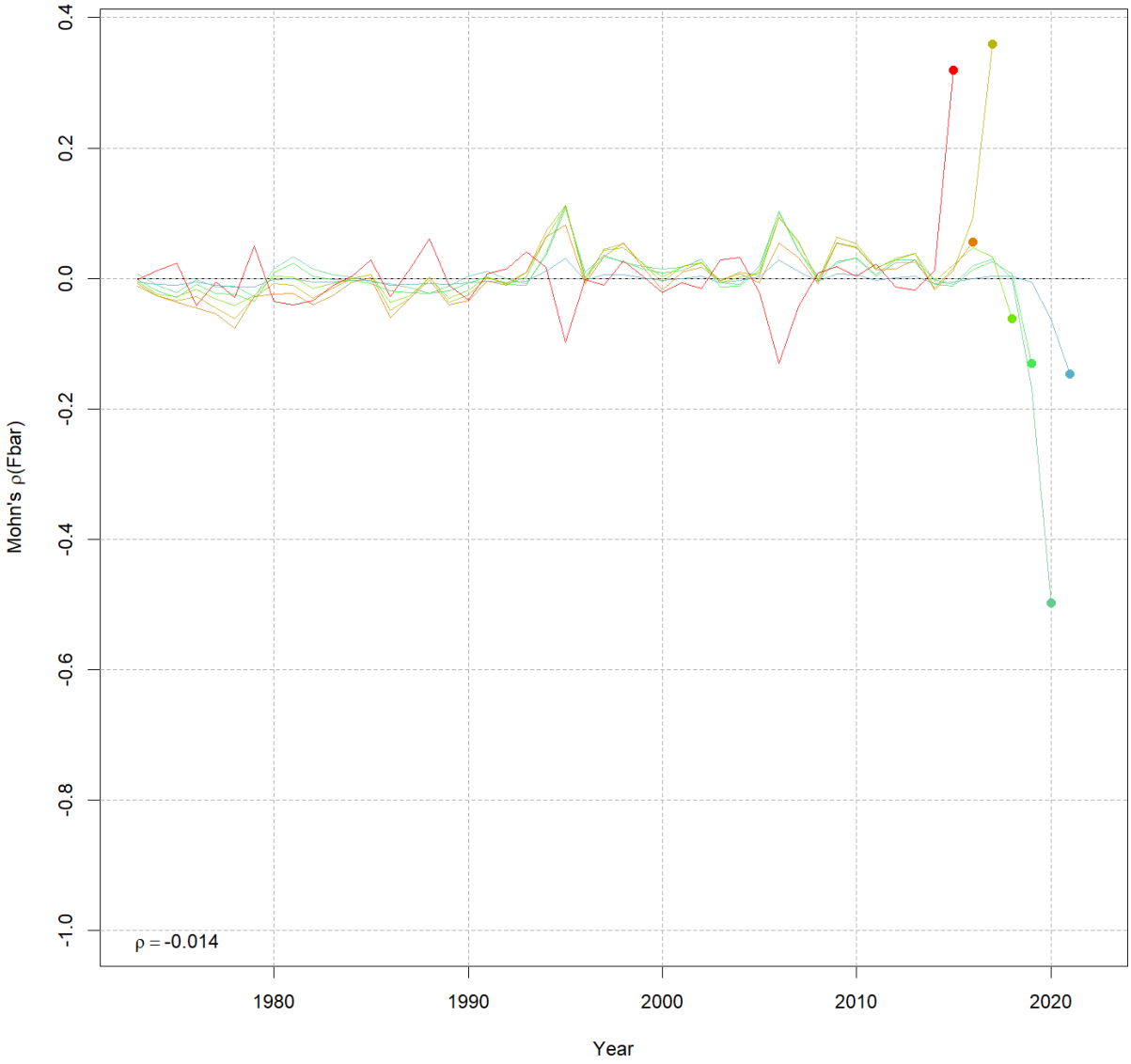

Figure 5.1.23. Relative retrospective pattern of F for the candidate model.

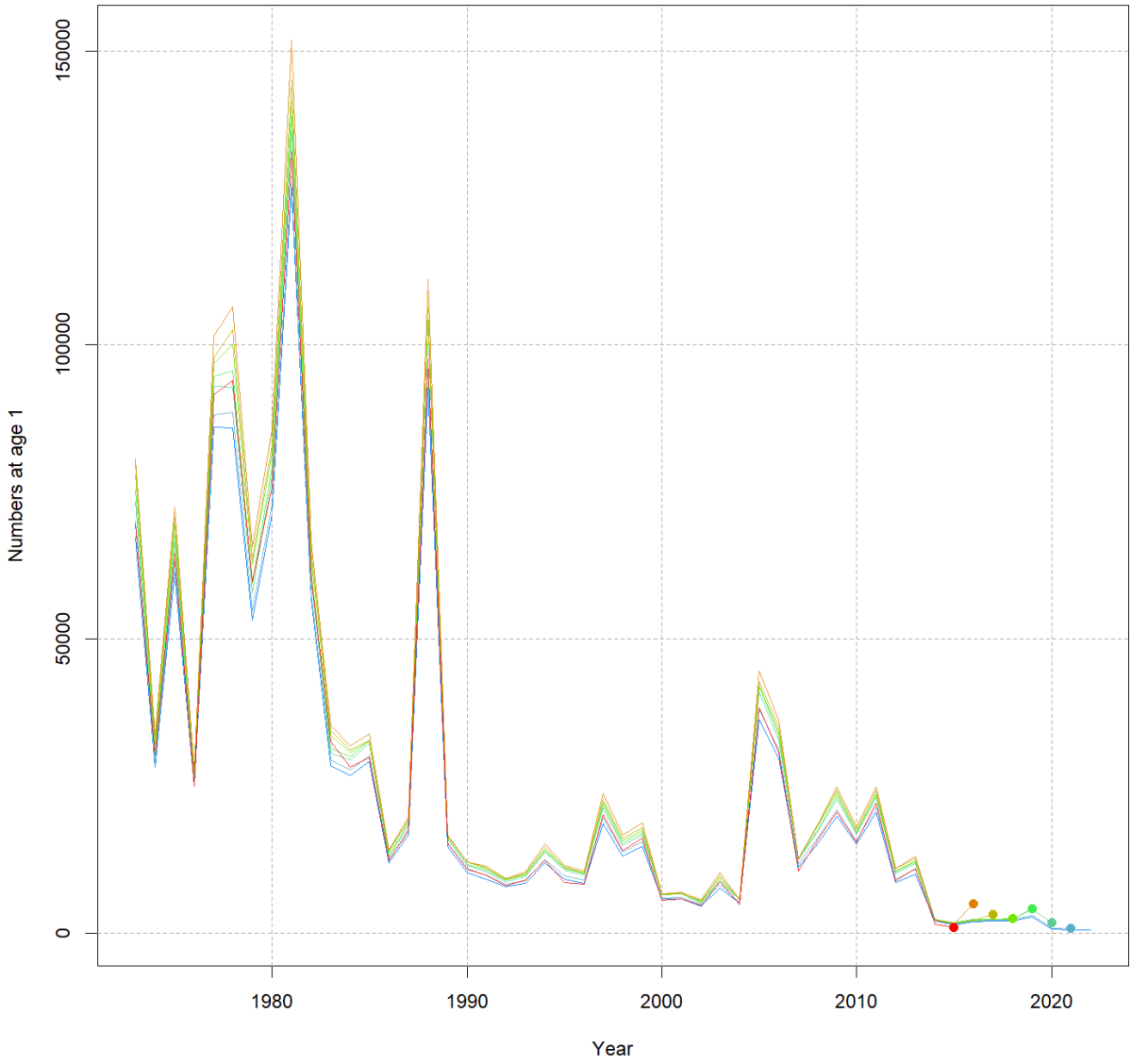

Figure 5.1.24. Retrospective pattern of R for the candidate model.

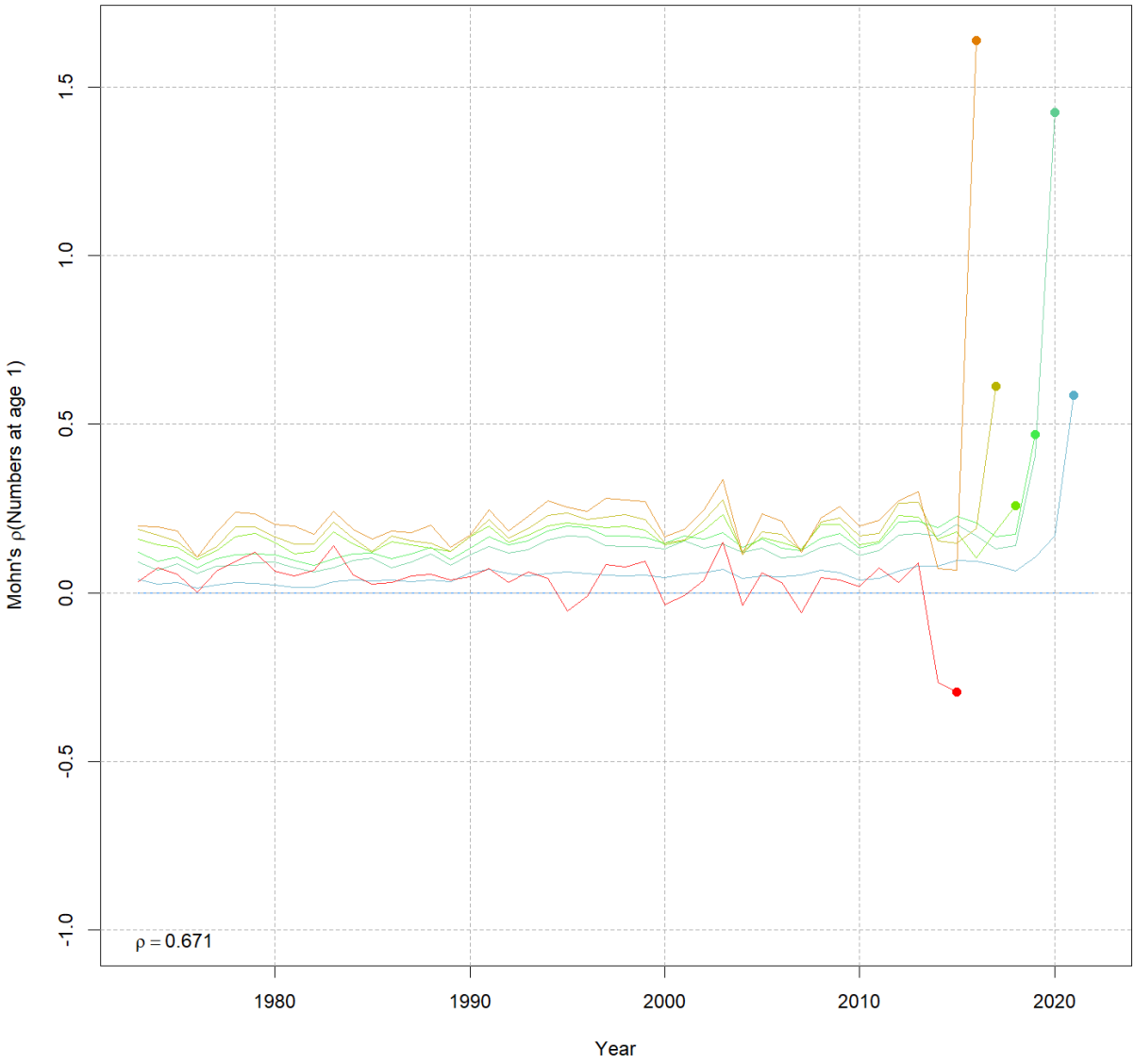

Figure 5.1.25. Relative retrospective pattern of R for the candidate model.

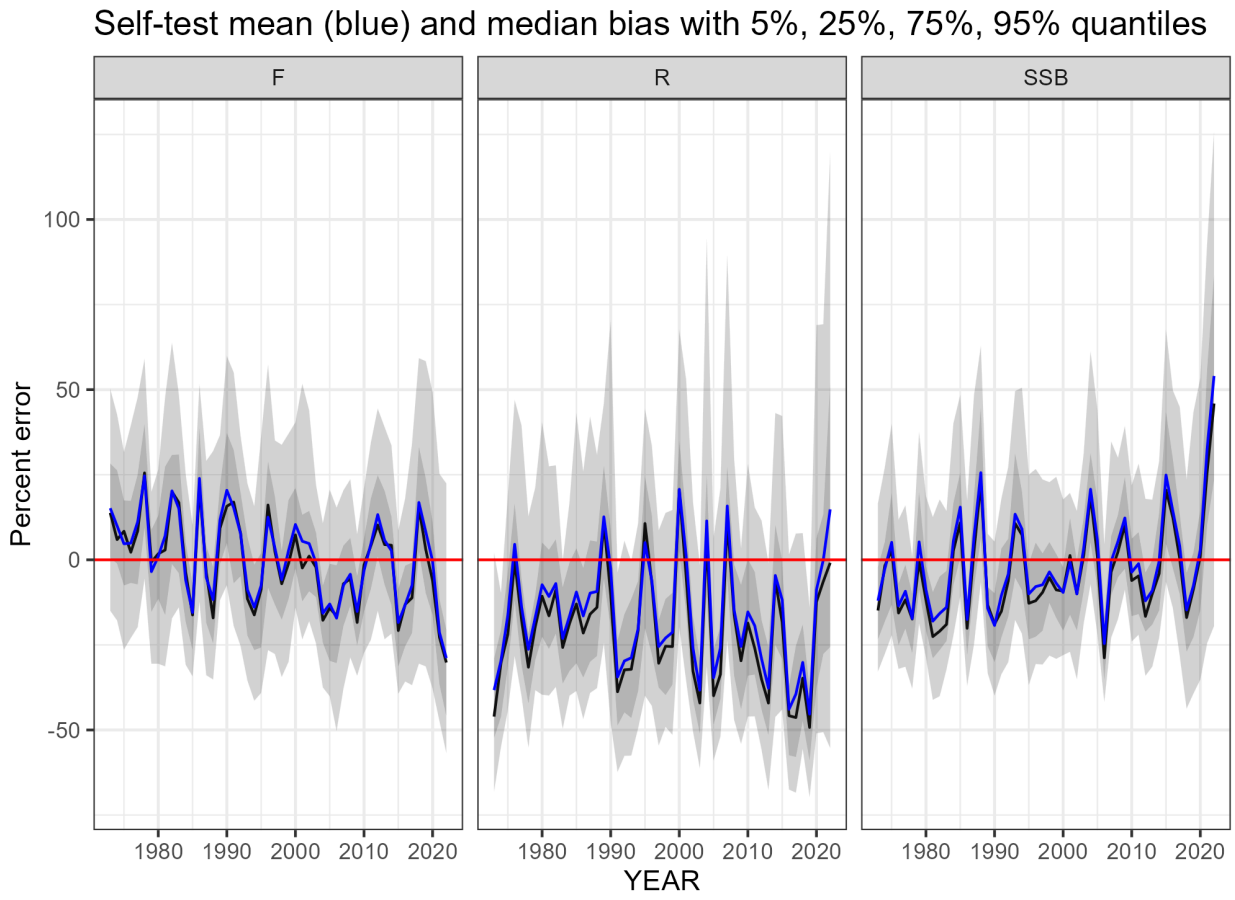

Figure 5.1.26. Self-test results of the candidate model.

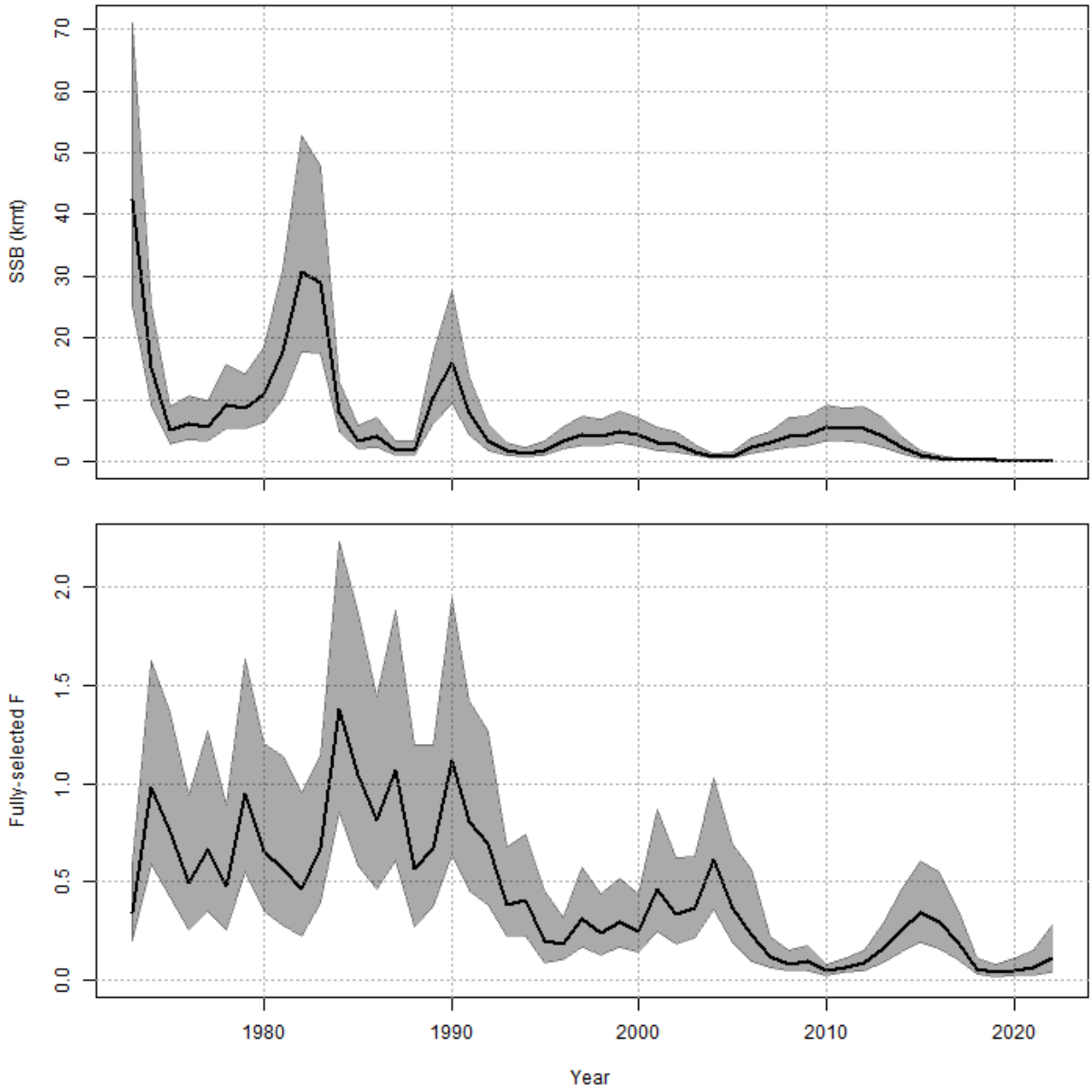

Figure 5.2.1. SSB (top) and F (bottom) from the candidate model.

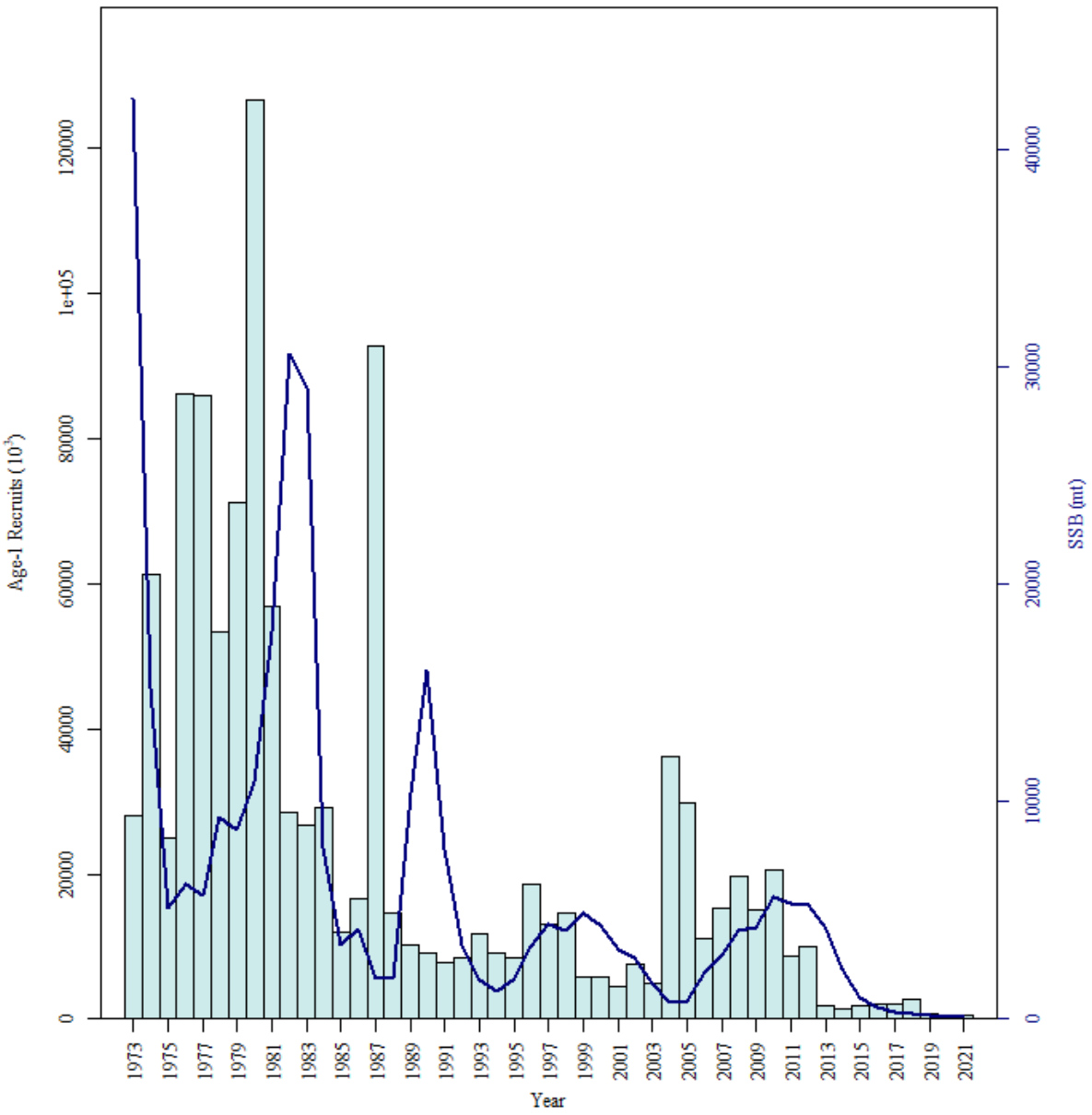

Figure 5.2.2. SSB (line) and R (bars) from the candidate model.

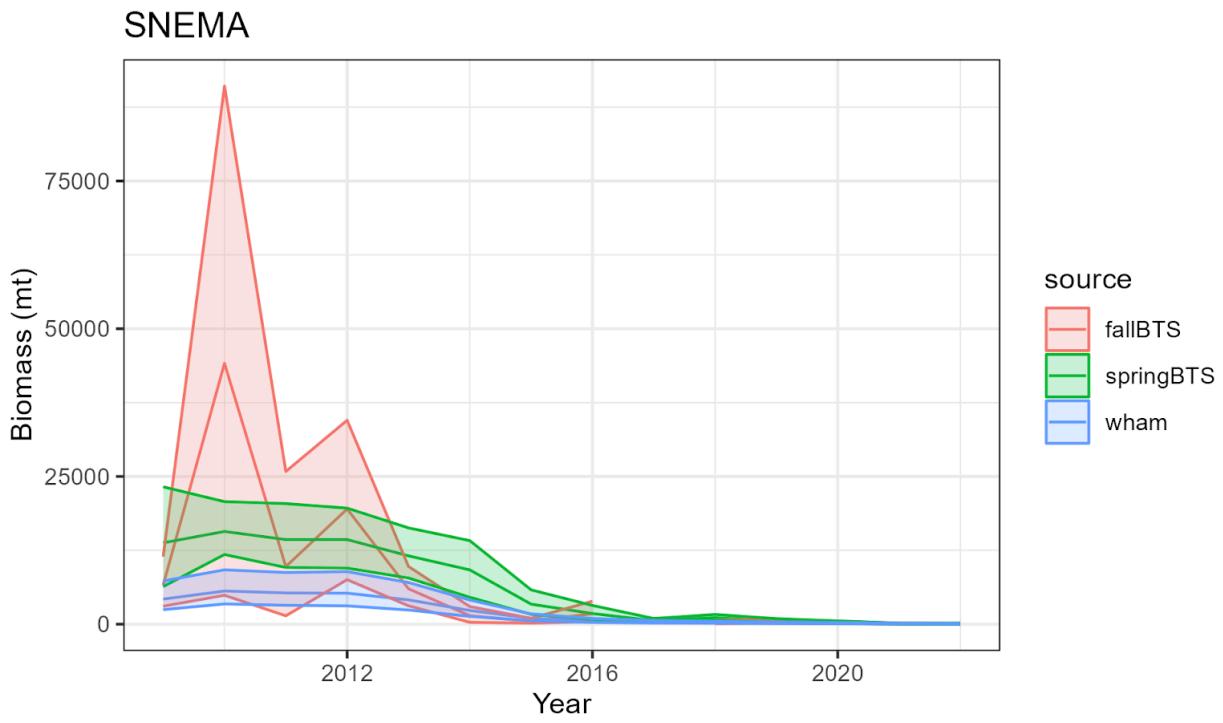

Figure 5.2.3. Chainsweep expanded survey biomass estimates 2009 to 2022 using data from the fall and spring NEFSC bottom trawl surveys (fallBTS and springBTS, respectively) compared to estimates from the candidate model using the WHAM framework (wham).

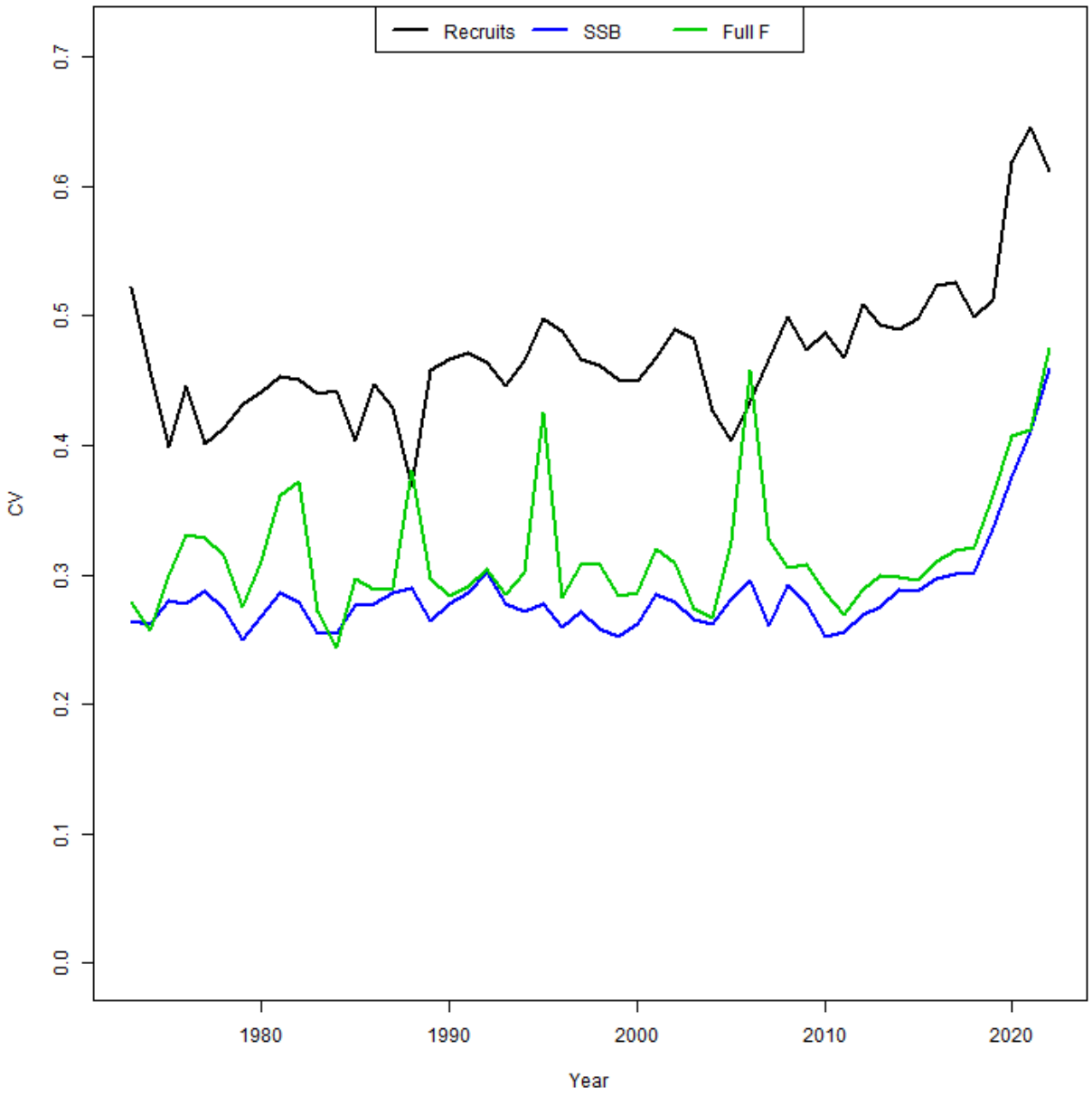

Figure 5.2.4. CVs of SSB, F, and R from the candidate model.

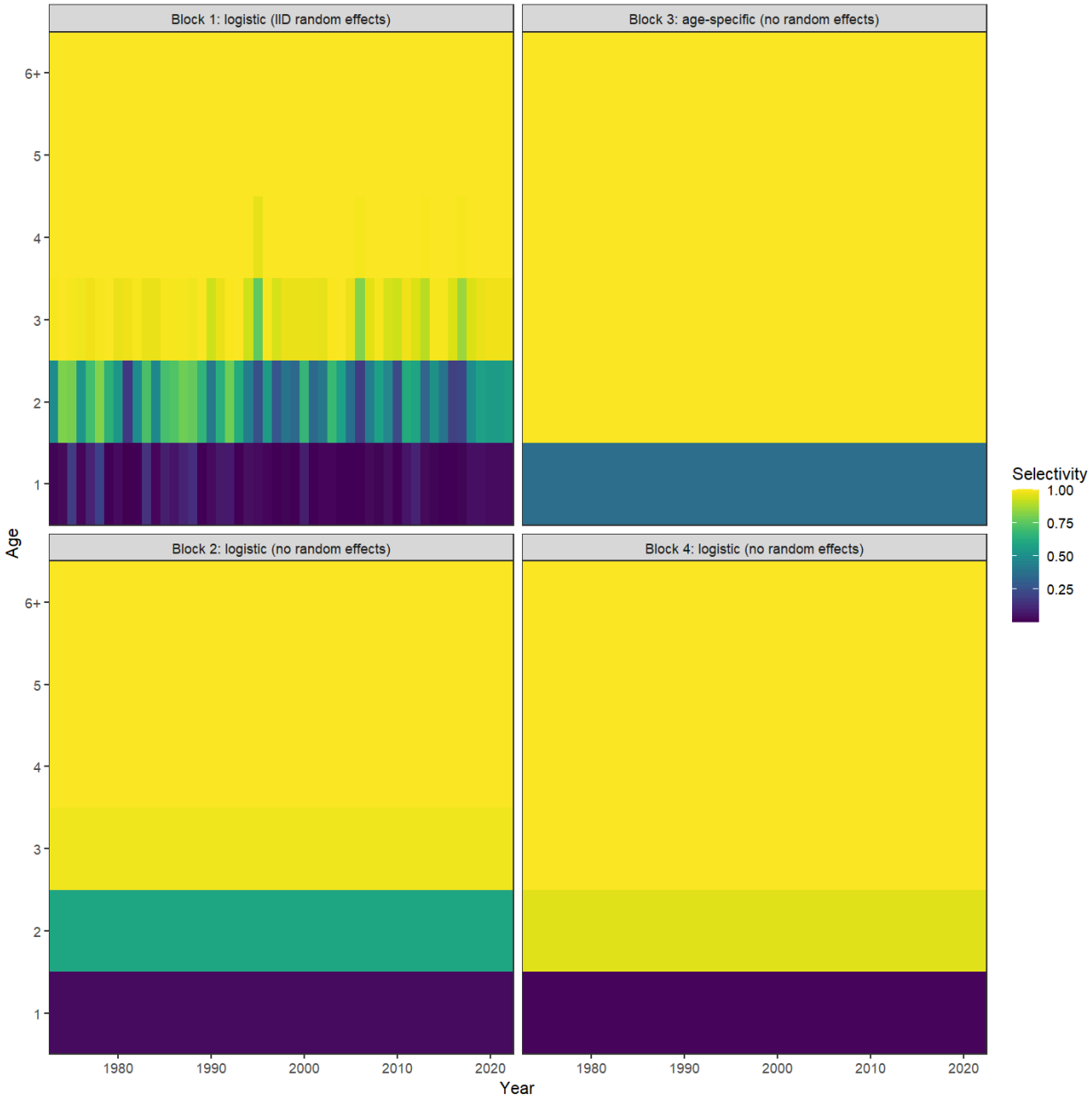

Figure 5.2.5. Aggregate fleet (top left), NEFSC spring (bottom left), NEFSC fall (top right), and NEFSC winter (bottom right) selectivities-at-age from the candidate model. Fleet selectivity changes across the time series 1973-2022 due to the iid random effects.

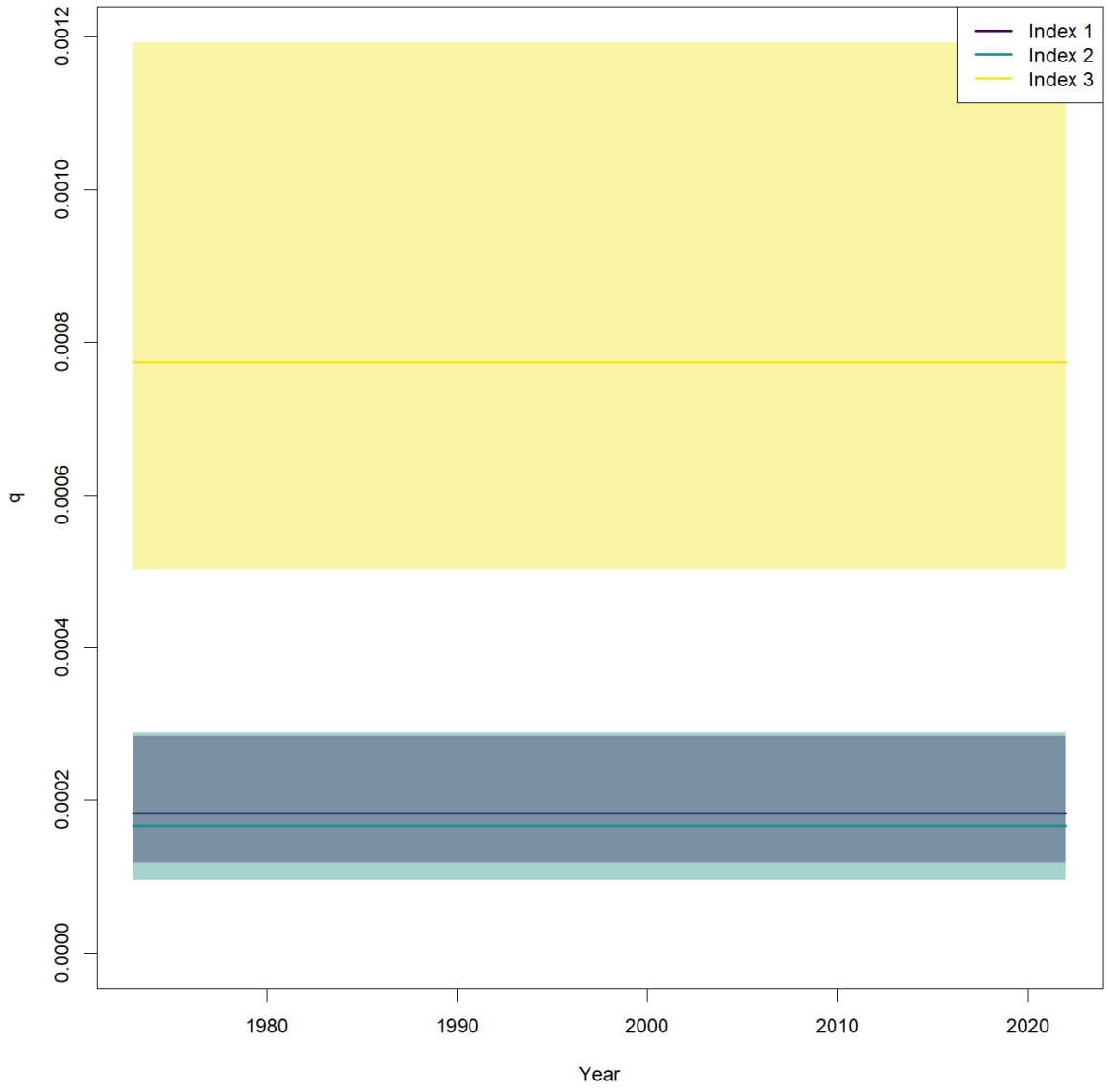

Figure 5.2.6. NEFSC spring (index 1), NEFSC fall (index 2), and NEFSC winter (index 3) catchabilities from the candidate model.

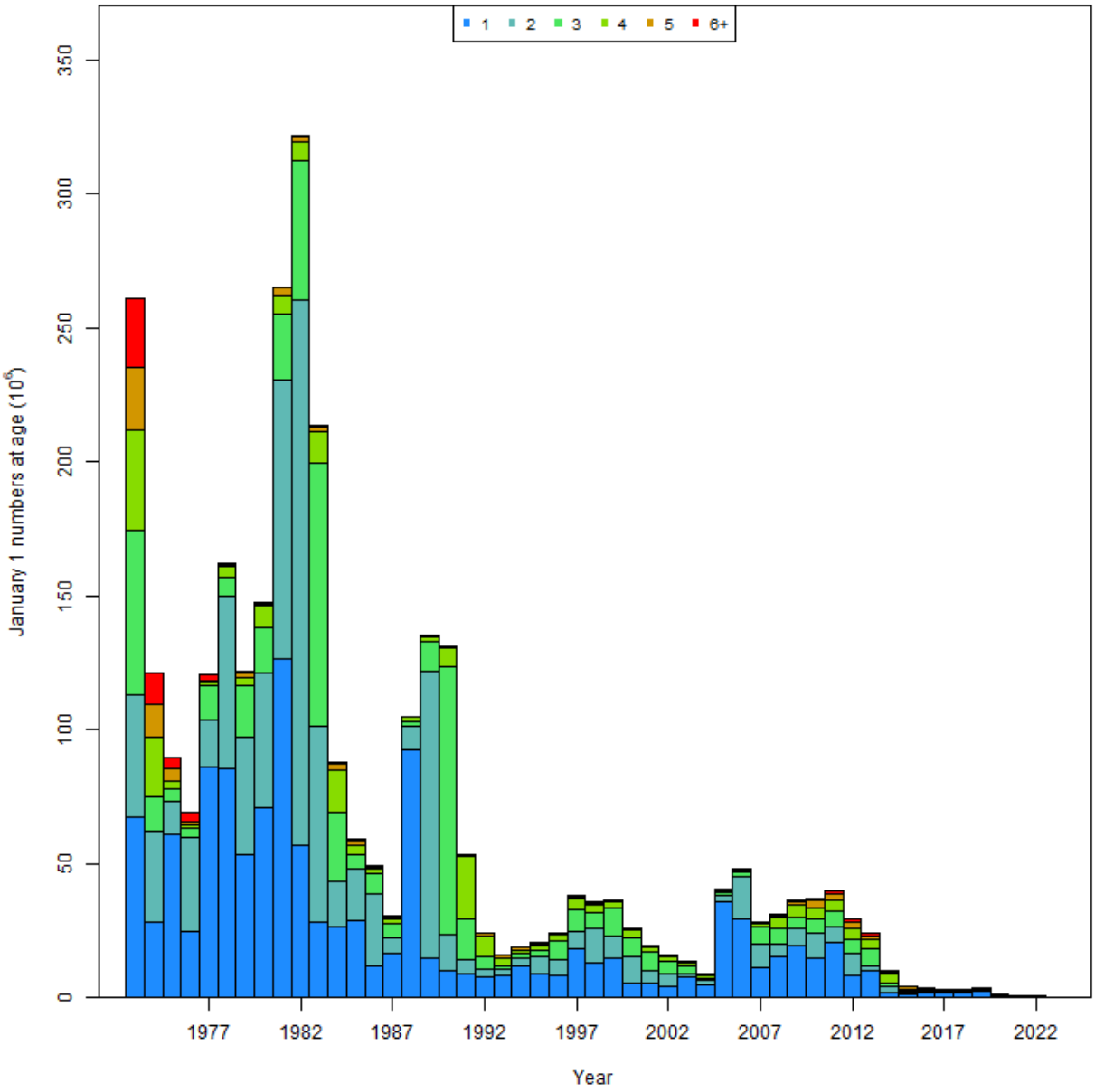

Figure 5.2.7. January 1st NAA from the candidate model.

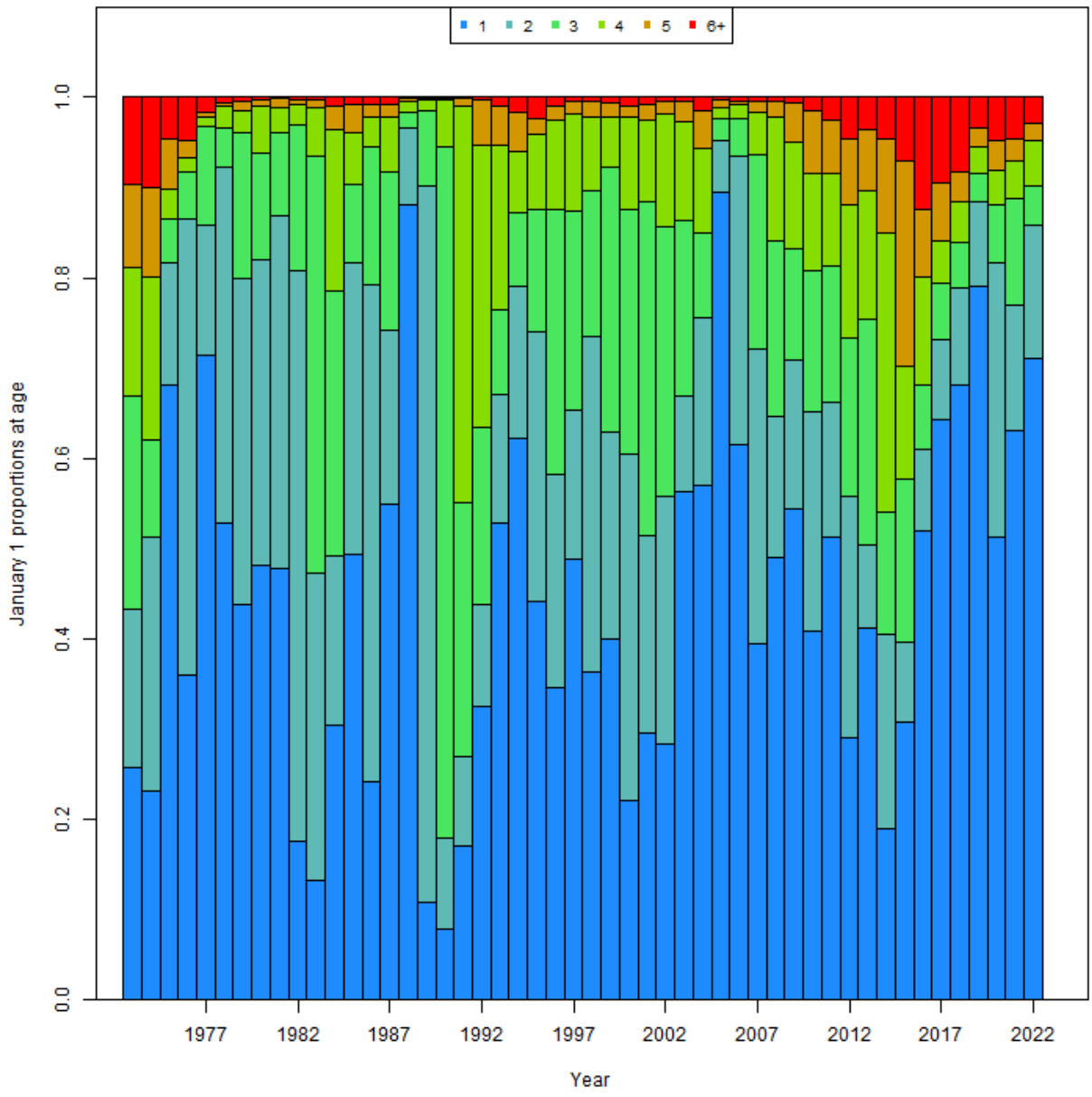

Figure 5.2.8. Proportional January 1st NAA from the candidate model.

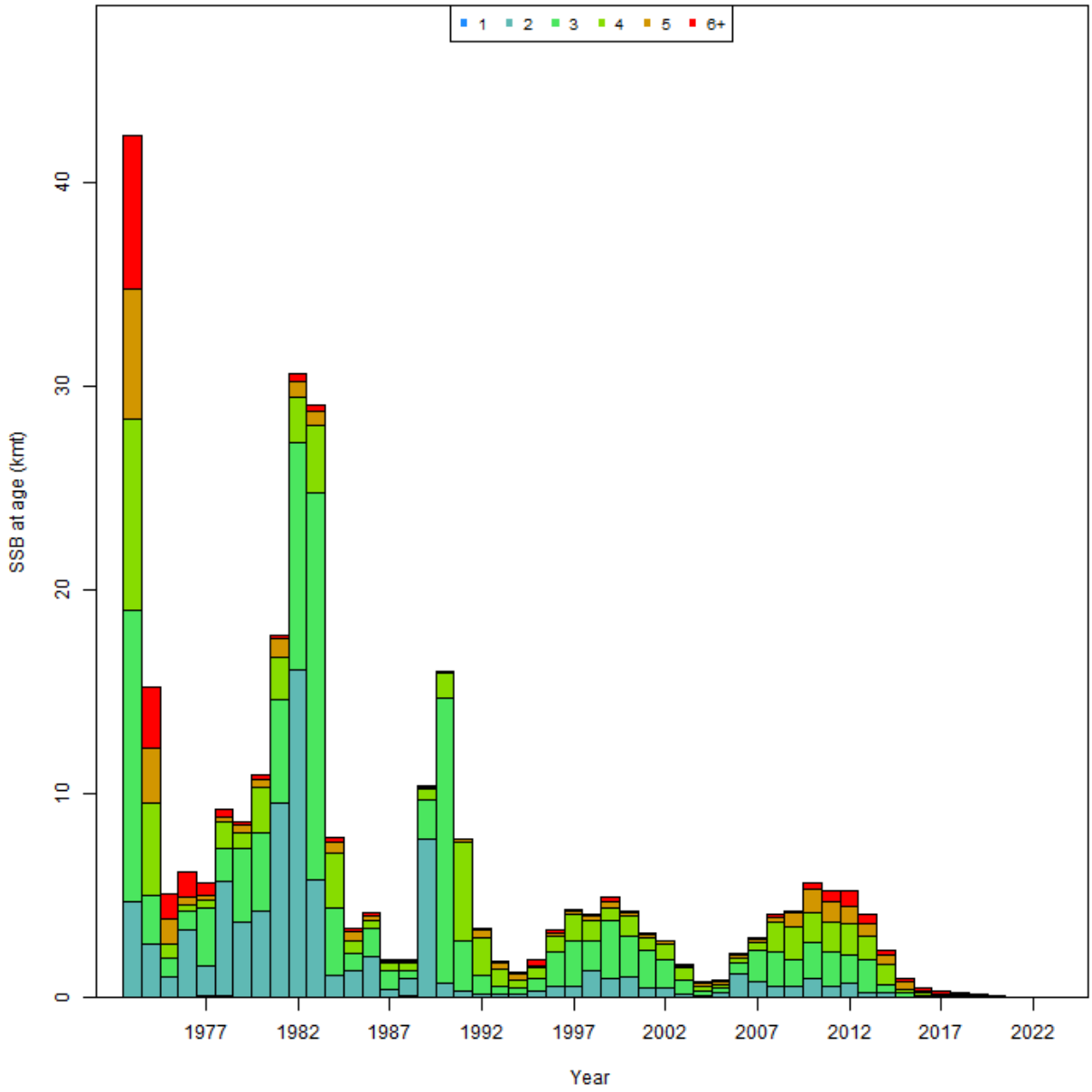

Figure 5.2.9. SSB-at-age from the candidate model.

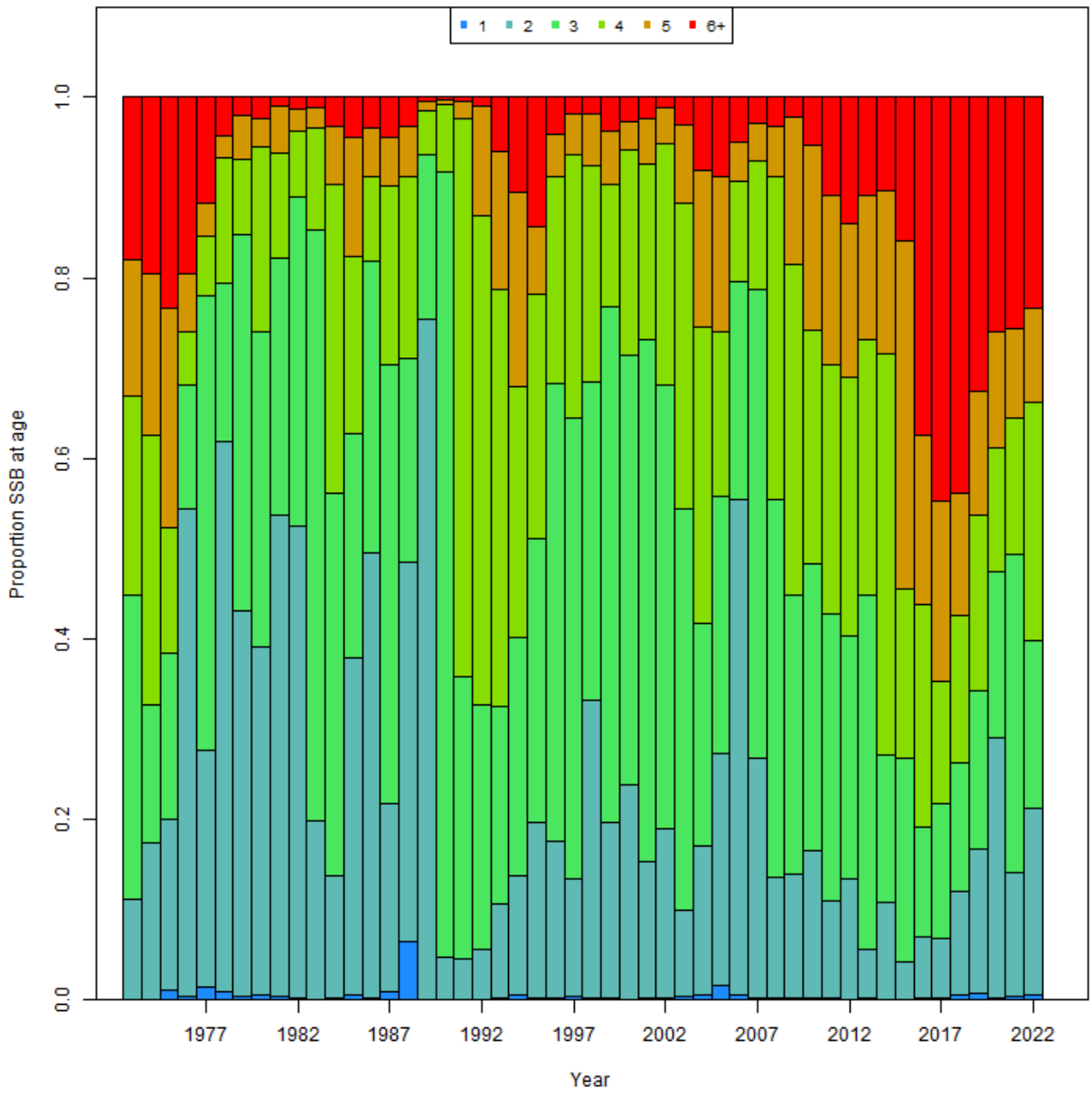

Figure 5.2.10. Proportional SSB-at-age from the candidate model.

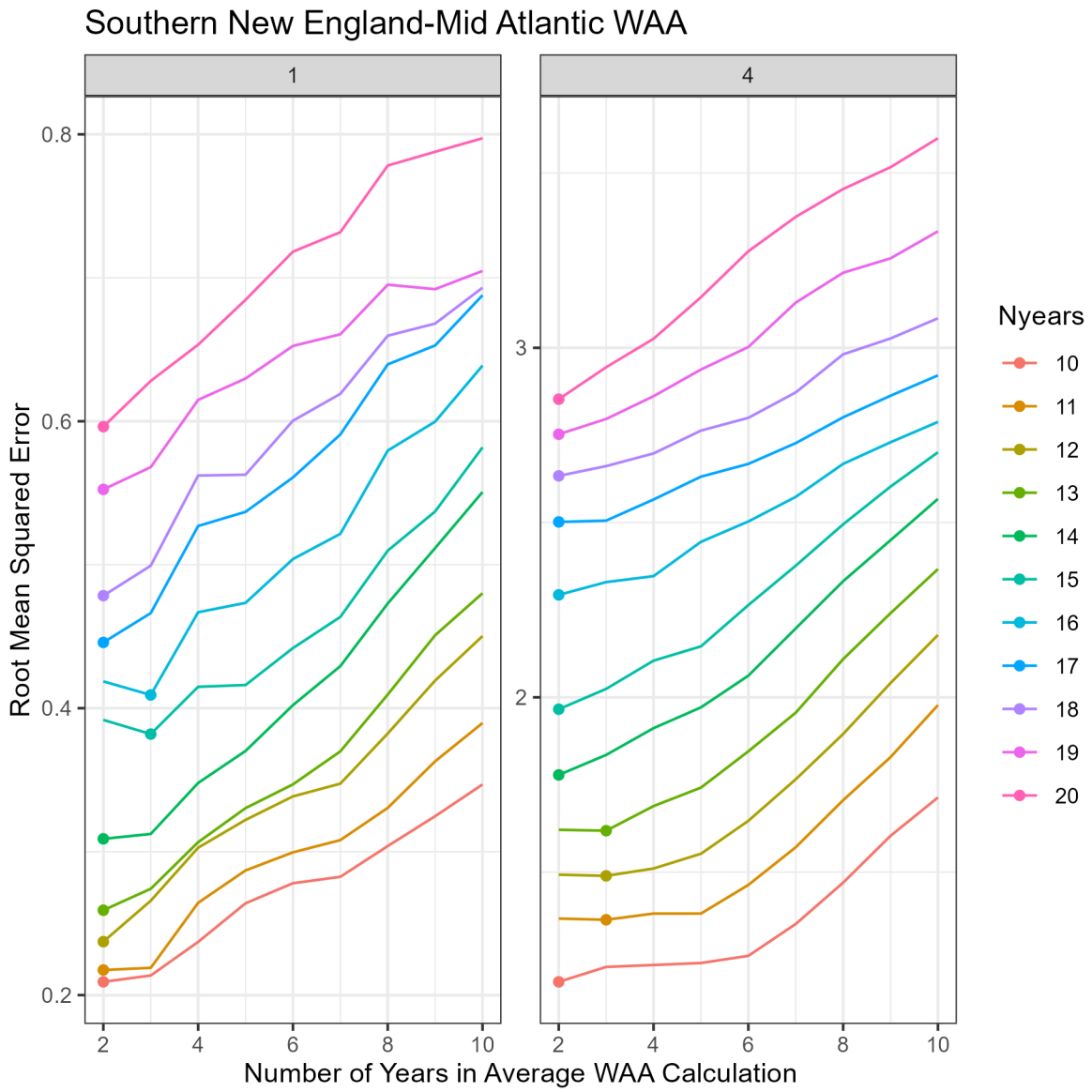

Figure 6.1.1. Moving window analysis for determining appropriate averages for WAA, Left is predicting one year ahead; right is predicting four years ahead. Colors represent the number of peels used to calculate the summed RMSE.

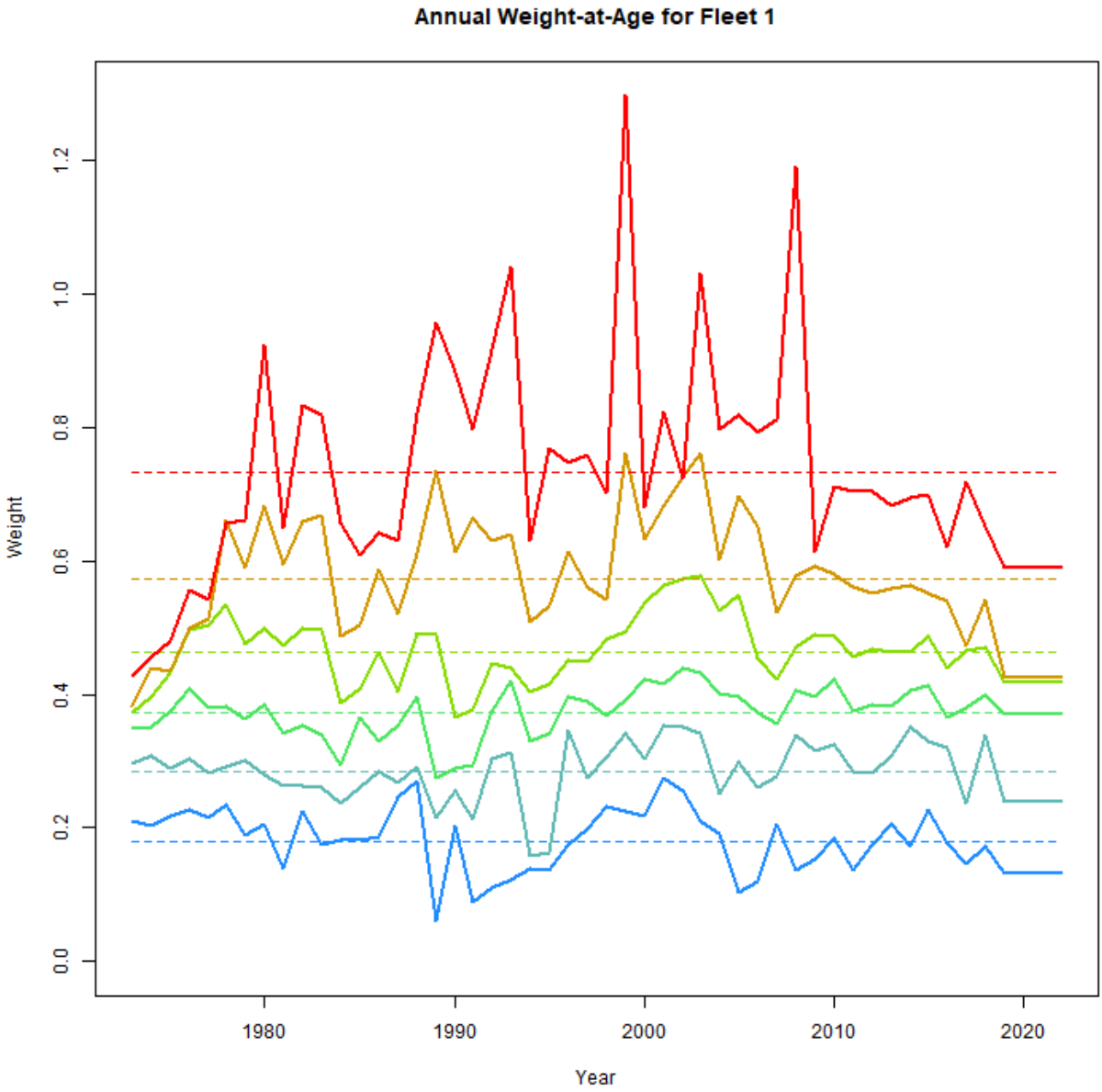

Figure 6.1.2. Aggregate fleet WAA over the series.

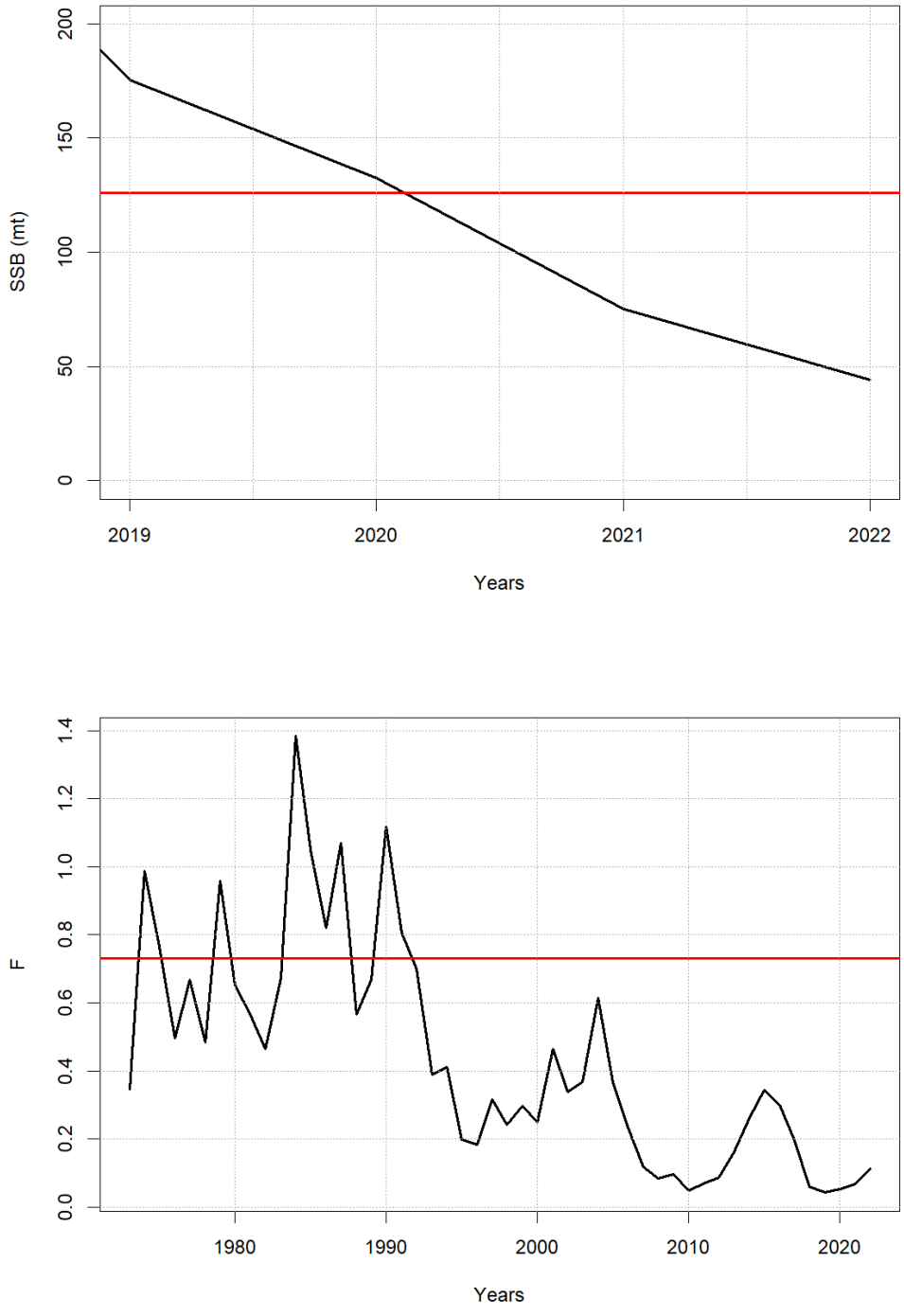

Figure 6.3.1. SSB 2019-2022 with a red line at $SSB_{40\%}$ = 126 mt (top) and F 1973-2022 with a red line at $F_{40\%}$ = 0.73 (bottom).

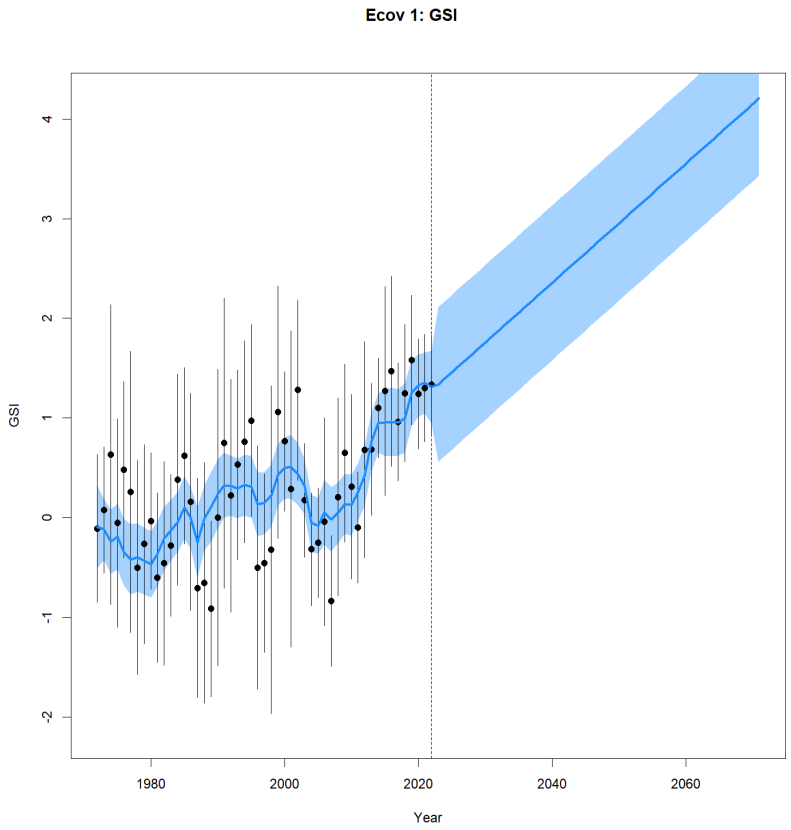

Figure 7.1.1. Long-term projection of GSI using a linear regression of years 2012-2022. Not used for candidate model projections.

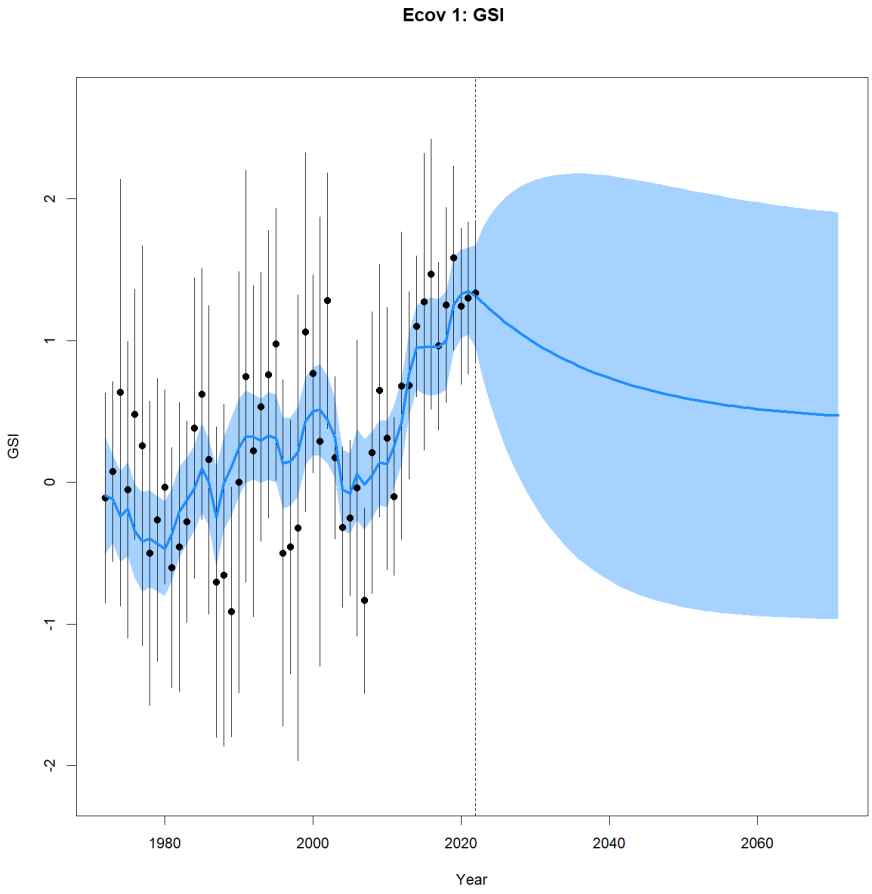

Figure 7.1.2. Long-term projection of GSI propagating forward the autoregressive (ar1) process. Not used for candidate model projections.

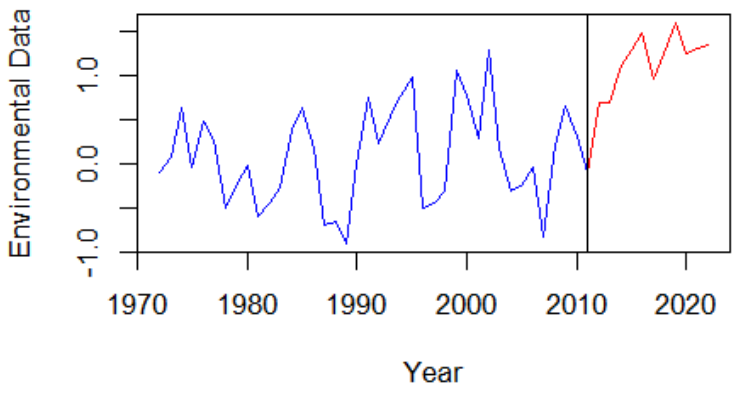

Figure 7.1.3. GSI 1973-2022. Blue and red mark the two series that produced the lowest absolute summed error in the changepoint analysis. Changepoint determined to be 2012.

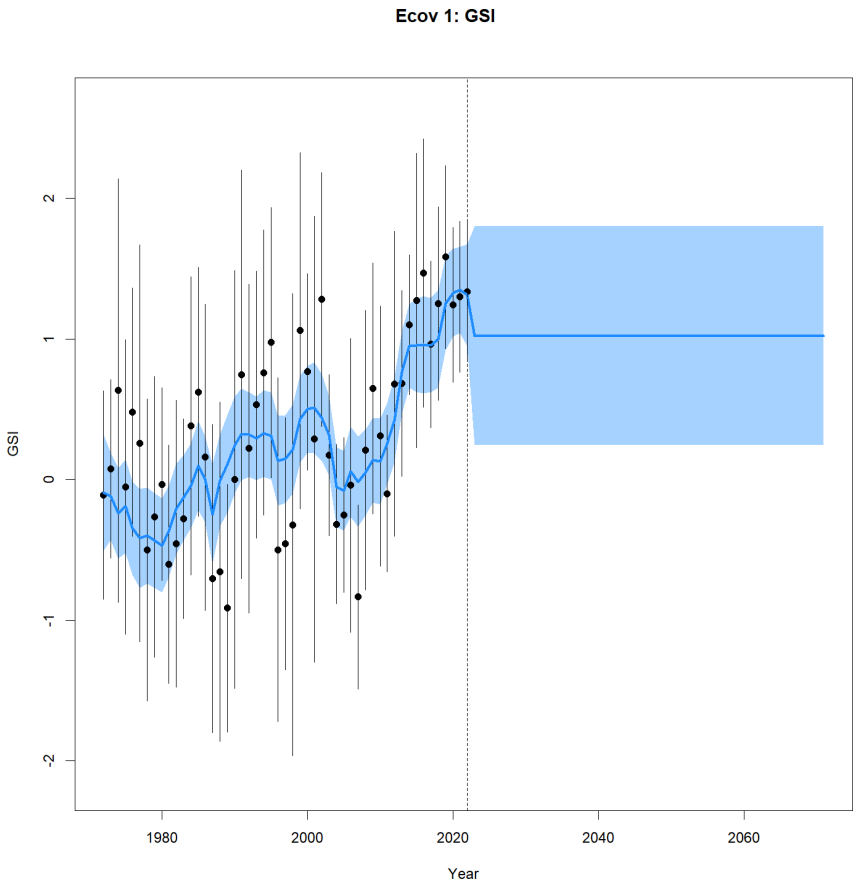

Figure 7.1.4. Long-term projection of GSI using a mean of years 2012-2022. Used for candidate model projections and reference point determination.

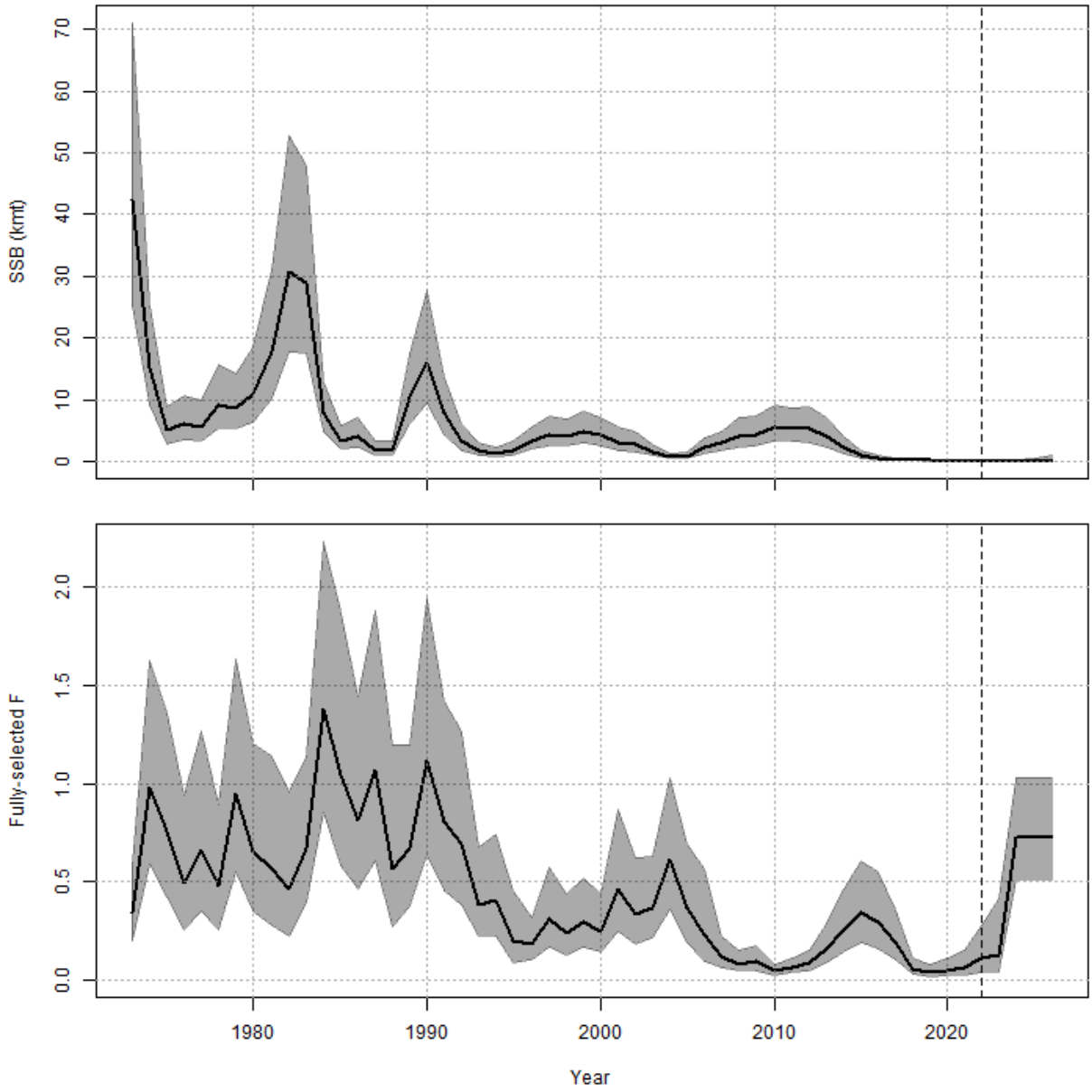

Figure 7.2.1. Short-term projections of SSB (top) and F (bottom).

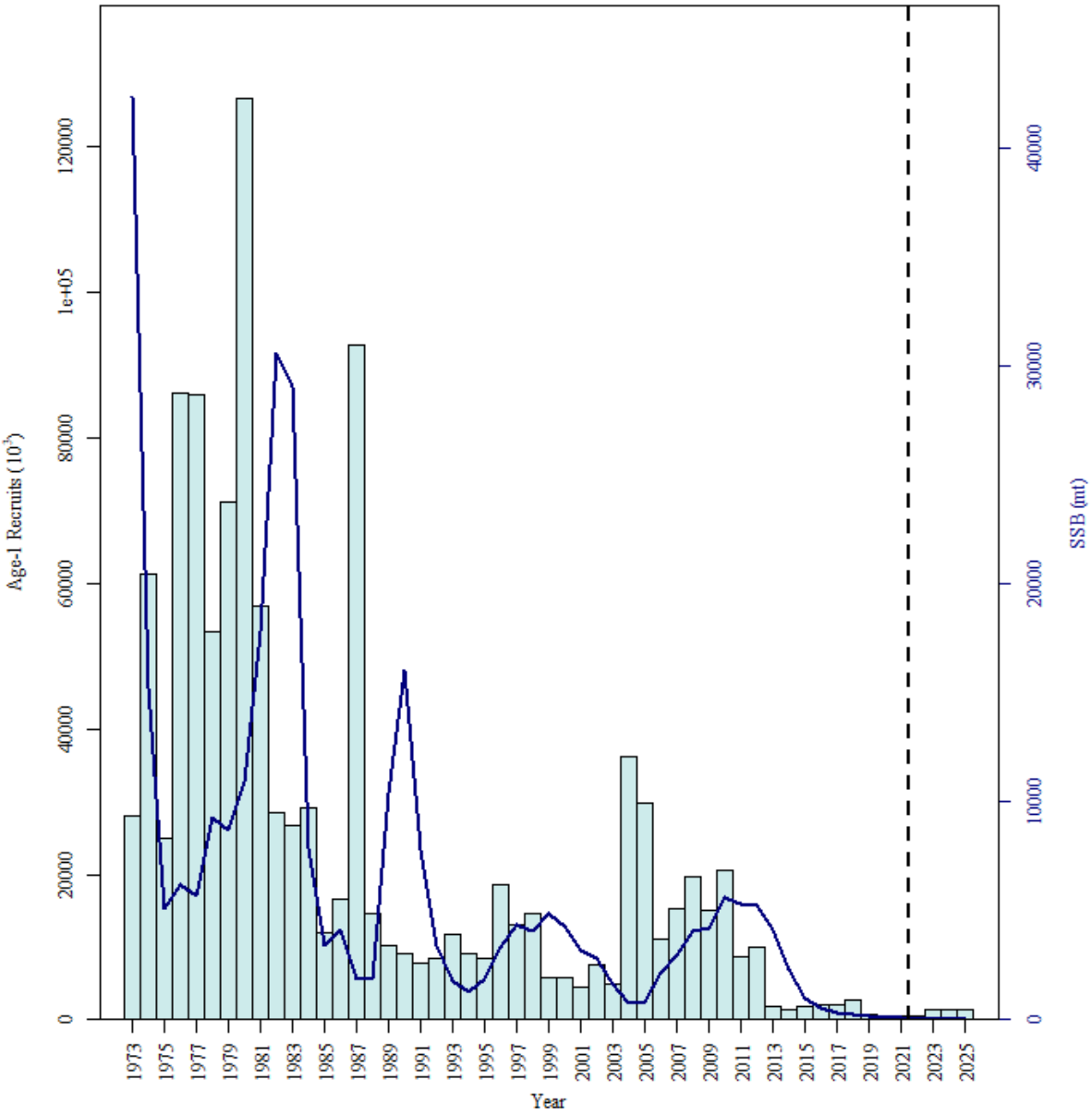

Figure 7.2.2. Short term projections of SSB (blue line) and recruitment (bars).

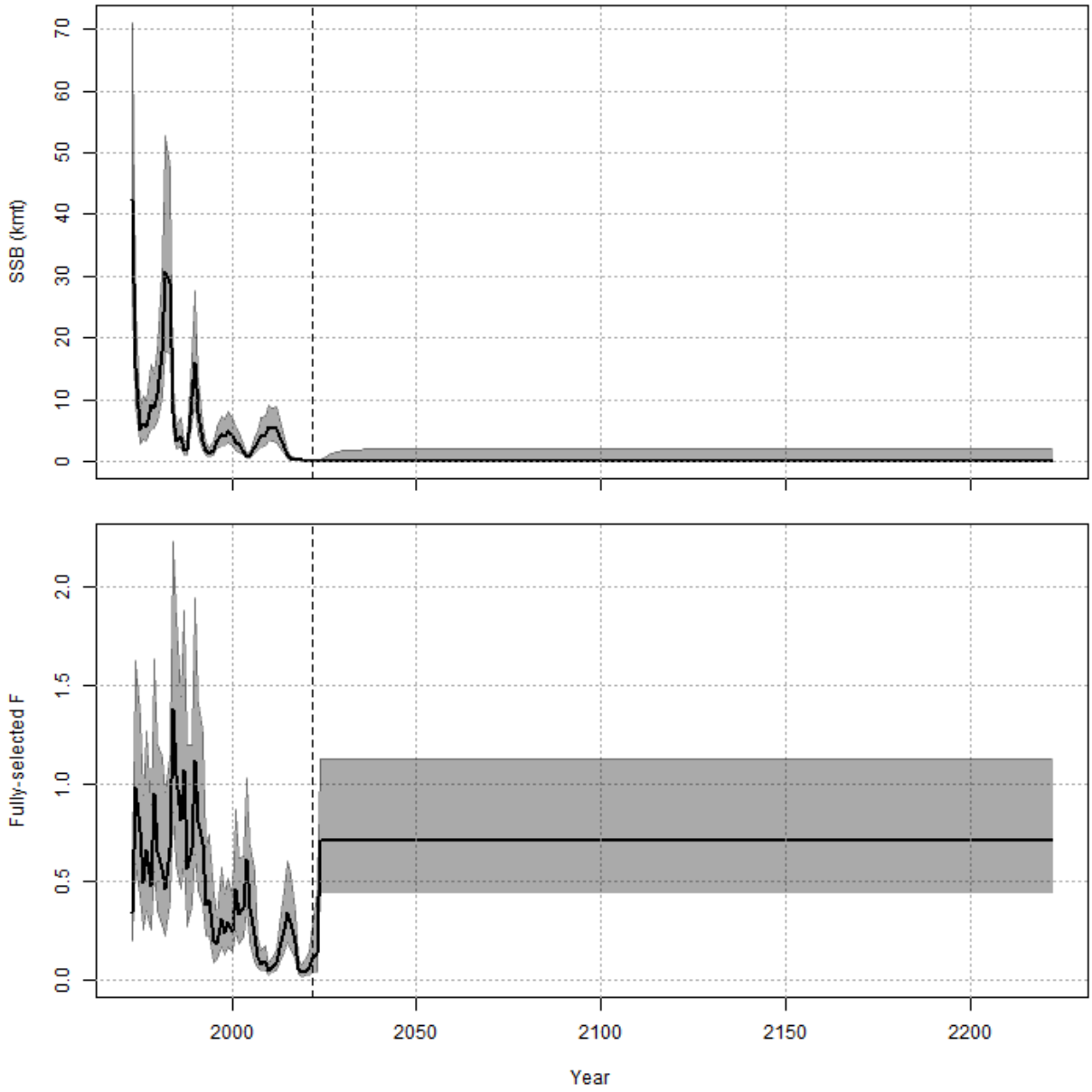

Figure 7.2.3. Long-term projections of SSB (top) and F (bottom).

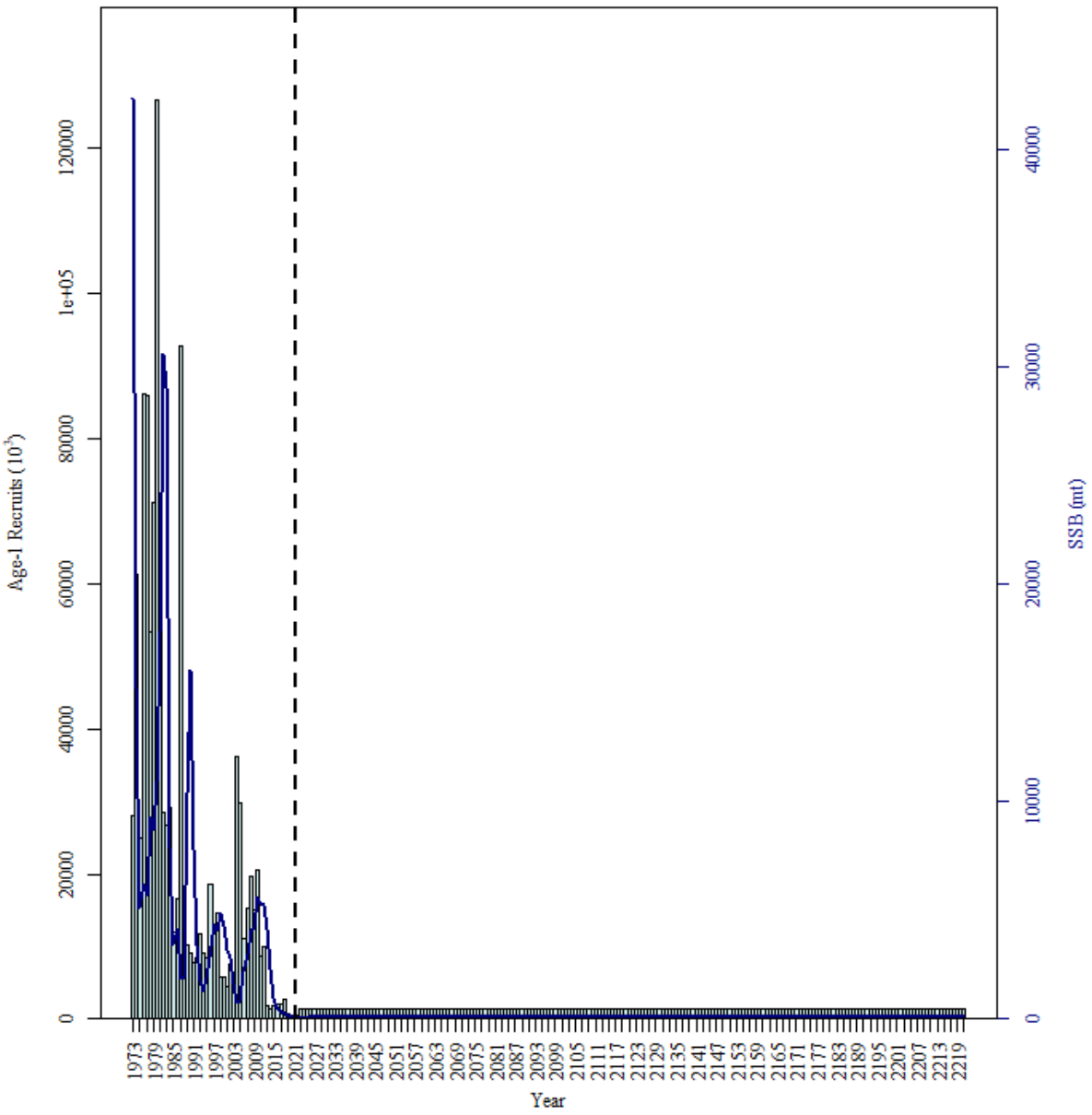

Figure 7.2.4. Short term projections of SSB (blue line) and recruitment (bars).

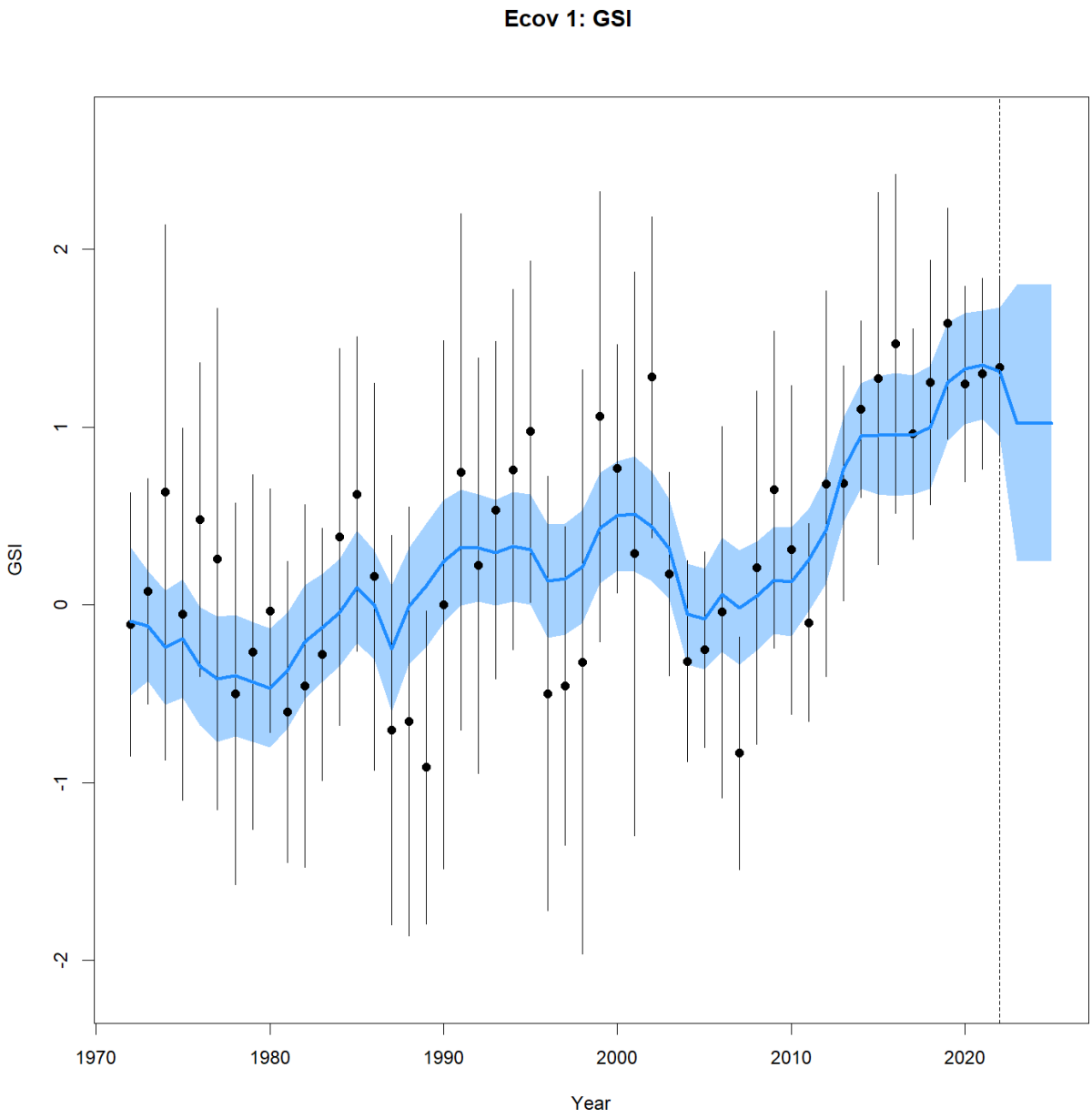

Figure 7.2.5. Short-term projection of GSI using a mean 2012-2022.