IPHC-2022-SRB020-07

Development of the 2022 Pacific halibut (*Hippoglossus stenolepis*) stock assessment

PREPARED BY: IPHC SECRETARIAT (I. STEWART AND A. HICKS; 11 MAY 2022)

Table of Contents

Summary

This document reports preliminary analyses in development of the 2022 Pacific halibut (*Hippoglossus stenolepis*) stock assessment. It follows the previous full stock assessment and independent peer review conducted in 2019 (Stewart and Hicks 2019b; Stewart and Hicks 2020; Stokes 2019), and subsequent updates to that assessment in 2020 (Stewart and Hicks 2021), and 2021(Stewart and Hicks 2022). Following the review of this document in June 2022 (SRB020), requested revisions will be considered and presented for additional review in September 2022 (SRB021), and the final 2022 assessment will be produced for the IPHC's Interim (IM098) and Annual (AM099) meetings. Updated data sources, including the results of the 2022 Fishery-Independent Setline Survey (FISS), logbook and biological data from the 2022 commercial fishery, and sex-ratio information from the 2021 commercial landings-at-age will be included for the final 2022 analysis.

Creating robust, stable, and well-performing stock assessment models for the Pacific halibut stock has historically proven to be challenging due to the highly dynamic nature of the biology, distribution, and fisheries (Stewart and Martell 2014). The stock assessment for Pacific halibut has evolved through many different modeling approaches over the last 30 years (Clark 2003). These changes have reflected improvements in fisheries analysis methods, changes in model assumptions, and responses to recurrent retrospective biases and other lack-of-fit metrics (Stewart and Martell 2014). The use of multiple models provides a solution to the endless search for a better stock assessment model and allows for structural as well as estimation uncertainty to be better captured. The IPHC adopted the ensemble approach for its 2012 stock assessment (Stewart et al. 2013a) and has continued to develop and refine the set of models used to provide tactical management information each year. The ensemble approach integrates the results of multiple hypotheses with the uncertainty associated with parameter estimation (Stewart and Martell 2015). This reduces potential for abrupt changes in management quantities as improvements and additional data are added to individual models (Stewart and Hicks 2018), and provides a more realistic perception of uncertainty than any single model, and therefore a stronger basis for probabilistic risk assessment.

Development of the current ensemble of stock assessment models began in 2012 with a single model using three alternative fixed values of natural mortality (Stewart et al. 2013a). In subsequent years, ensemble development included exploration of highly varied model approaches, including a Virtual Population Analysis (VPA) and a simple biomass production model (Cox et al. 2014) and a spatially explicit model including migration rates and recruitment distribution (Cox et al. 2017). The treatment of the historical data through long and short modelled time-series, and the treatment of spatial patterns via coastwide aggregation of data and an Areas-As-Fleets (AAF) approach have emerged as two critically important axes over which to describe the uncertainty in both the scale and trends of the Pacific halibut stock and population dynamics. Therefore, recent ensembles have included four equally weighted models representing a two-way cross of time-series length (short and long) and data aggregation (coastwide and by Biological Region).

Starting with the 2021 stock assessment data, models and results (Stewart and Webster 2022; Stewart and Hicks 2022), this analysis provides a sequentially updated ‘bridge’ of the changes made thus far toward a preliminary assessment for 2022. This bridging analysis included a series of steps for which intermediate results and comparisons are provided. These steps included:

1) Extending the time series to include projected mortality based on limits adopted for 2022 (IPHC 2022),
2) updating to the newest stock synthesis software version (3.30.19; Methot Jr et al. 2021a),
3) expanding the treatment of natural mortality (*M*) to include an informative prior and increased values at the youngest ages based on meta-analyses,
4) improving the basis for data weighting via use of bootstrapped effective sample sizes based on the FISS and fishery sampling programs as model inputs (rather than the raw number of sets/trips),
5) re-tuning the process and observation error components of these models to achieve internal consistency within each,
6) allowing for interannual variability in the sex-ratio of the commercial fishery selectivity,
7) and exploring whether female *M* in the short models was estimable (male *M* and *M* for both sexes in the long models was already estimated).

Briefly, software versions, use of a prior on *M* and age-specific *M* for the youngest ages had little to no effect on individual model results. Time varying sex-ratio in selectivity for the commercial fishery and *M* in the short AAF model were both found to be robustly estimated. Retuning the sample sizes and process error variance terms provided for internal model consistency, and effects on results were similar to those in previous assessments. Convergence, sensitivity and retrospective analyses were performed on all models contributing to the ensemble. Alternatives to the treatment of the PDO as a covariate to average recruitment (long models only) were explored, but none were found that outperformed the *status quo*. All models were sensitive to the estimated or fixed value of female *M*, with increasing *M* always resulting in larger estimates of spawning biomass. After including time-varying sex-ratio of the commercial fishery selectivity, retrospective analyses were much more stable than in previous assessments and showed little trend as data were removed. Jitter analyses indicate that the long AAF model was the least robust to a wide range of initial parameter estimates; however, convergence did appear to be achieved.

After evaluating individual models, the analysis also included an exploration of model weighting within the ensemble; models have been equally weighted since the 2013 stock assessment. The Mean Absolute Scaled Error (MASE; Hyndman and Koehler 2006) of one-year-ahead projections of the FISS coastwide index of abundance suggested that all four preliminary models performed appreciably better than the naïve projection (last year’s index). When this performance was used to weight the models, weights ranged from 9 to 38% across a 1-4 year historical window. The highest weights were generally assigned to the coastwide long model, and the lowest to the AAF short model. A MASE-weighting approach would provide a self-updating approach for model weights within the ensemble that is logically linked to the prediction skill of the quantity most relevant to management decision-making and is proposed for use in the final 2022 assessment. In aggregate, the results of the preliminary ensemble across a range of individual model weights remain consistent with those from recent assessments. The uncertainty in stock dynamics also remains similar and high relative to that frequently reported for many single-model or simple stock assessment analyses. This uncertainty will continue to be captured via the annual decision table (Stewart and Hicks 2022), reporting the trade-offs between yield and various stock and fishery risks.

Given the challenges and uncertainties of the Pacific halibut population dynamics and stock assessment it is unlikely that future assessment models will provide substantially more precise and stable results, even as data time-series grow longer. In light of the uncertainty and variability within which the Pacific halibut management occurs, a robust management procedure, tested via the IPHC's Management Strategy Evaluation (MSE) process (Hicks and Stewart 2022) may provide a stronger basis for future management success and stability than annual decisions based on stock assessment results.

Data sources

The Pacific halibut data sources are collected with sampling designs created to produce results first for each IPHC Regulatory Area, and then to be aggregated to Biological Regions and to the entire range of the species in U.S. and Canadian waters (Figure 1). This section provides a brief overview of the key types of data available for analysis. A more in-depth summary can be found in the annual overview of data sources created each year and most recently for the 2018 stock assessment (Stewart and Webster 2019). Where specific improvements to existing data sources have been included in this assessment (i.e., sex-ratios from the 2017 commercial landings and the revised modelled survey time-series) changes are described below.

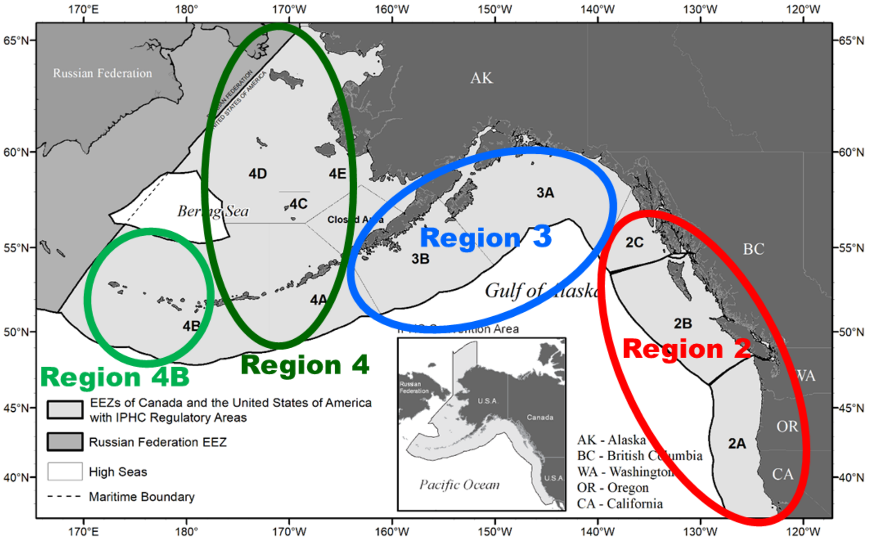


Figure 1. IPHC Regulatory Areas, Biological Regions, and the Pacific halibut geographical range within the territorial waters of Canada and the United States of America.

Overview of existing data

The time-series' of Pacific halibut data (described and plotted in much more detail in Stewart and Webster 2022) provide a rich historical record including mortality estimates, abundance indices (Catch-Per-Unit-Effort; CPUE) and age-composition data that extend back to the late 1800s and early 1900s (Figure 2). The IPHC's Fishery Independent Setline Survey (Ualesi et al. 2022; Webster 2022) provides the primary index of abundance and the most rich source of demographic information via individual weight, length and age data. The FISS includes Pacific halibut as young as 4-5 years old, which are below the IPHC's 32 inch (82 cm) minimum size limit (Stewart et al. 2021). Thus, these fish are observed several years prior to entry into the retained fishery landings which are sampled at the point of landing (Kong et al. 2022) and do not contain biological or catch-rate information on younger fish. Annual mortality estimates are provided to the IPHC from a variety of sources (Kong et al. 2022) including the directed halibut fisheries (commercial, recreational and subsistence) as well as incidental mortality associated with discards in directed fisheries and discard mortality in non-directed fisheries ('bycatch') that are not allowed to legally retain Pacific halibut. Each of these sources have differing levels of precision and likely accuracy associated with the estimates used for stock assessment.

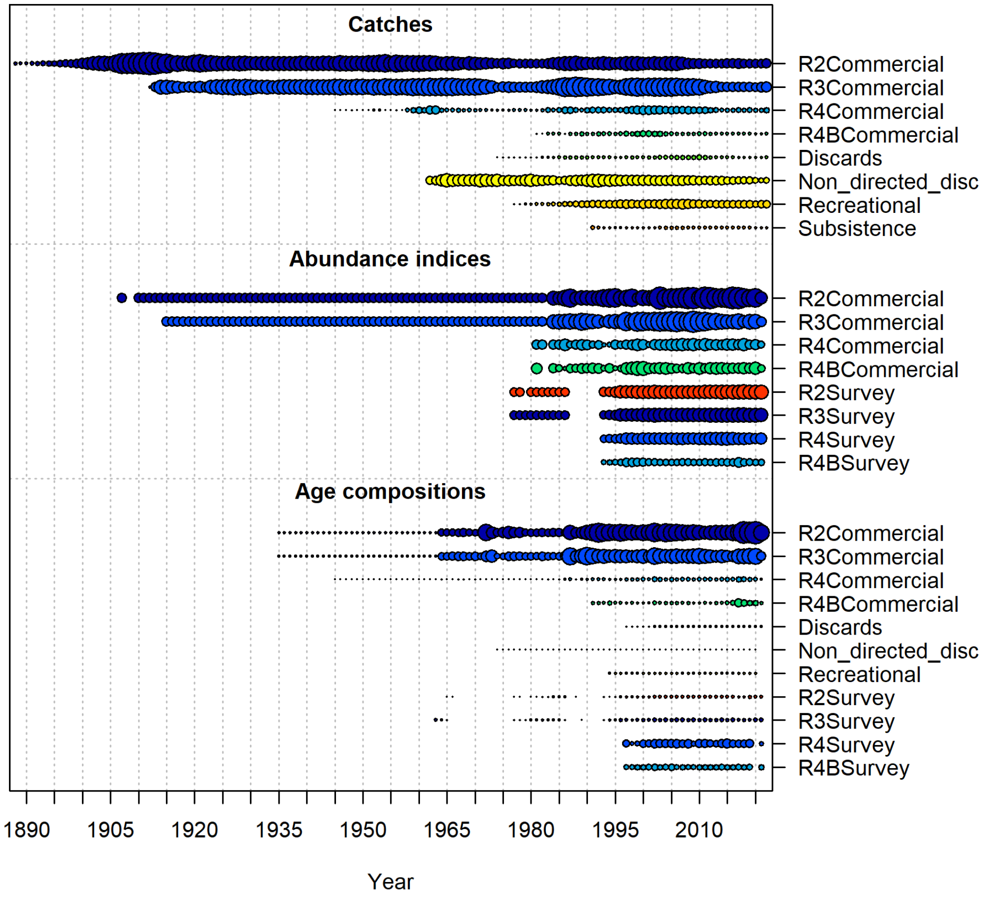


Figure 2. Data used in the stock assessment. Circle size is proportional to the magnitude of mortality (catches), inversely proportional to the variance (abundance indices) or proportional to the sample size (age-composition data).

Mortality

The industrial Pacific halibut fishery developed first off the west coast of the United States and Canada and sequentially moved to the north (Stewart and Webster 2022), only reaching full exploitation across all spatial areas in the last several decades. Mortality from non-directed discards increased rapidly with the arrival of foreign fleets into U.S. and Canadian waters in the 1960s. Recreational mortality has also increased over the time-series, although somewhat more gradually, since its initiation in the 1970s (Figure 3).

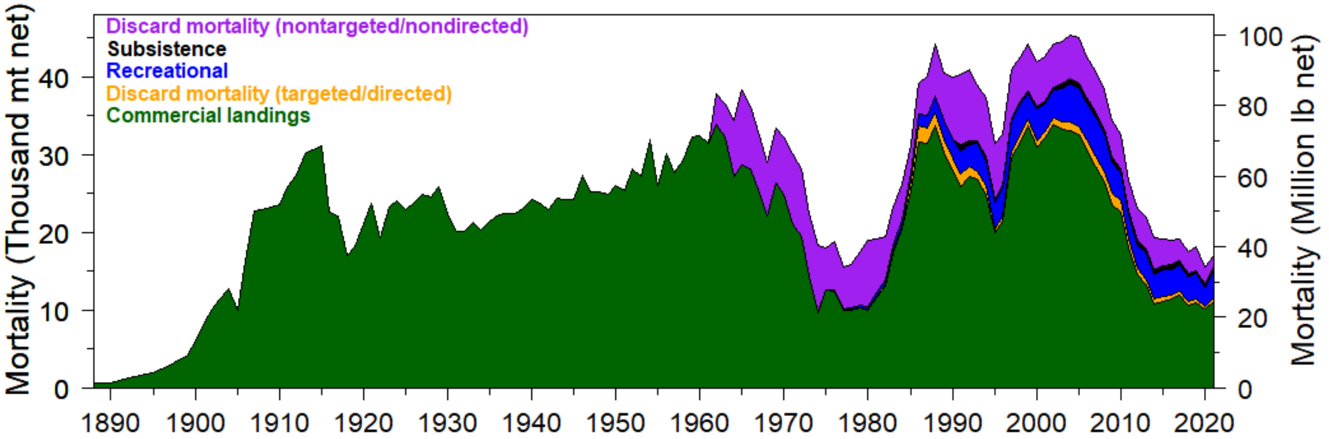


Figure 3. Time-series of mortality estimates by source.

Index data

The IPHC's FISS comprises the primary index of recent abundance and the primary source of biological data for use in the stock assessment. Index values (Table 1) are used in this assessment in numbers of halibut captured per unit effort (NPUE). The recent time-series (1993-2021) is based on the output of the IPHC's space-time model (Webster 2022; Webster et al. 2020) which estimates the degree of spatial and temporal correlation among survey stations in order to predict trends in biomass and abundance across the entire range of Pacific halibut within the IPHC Convention Area. This index provides precise trend information by IPHC Regulatory Area, which are weighted by the relative spatial bottom area and combined to Biological Regions and a coastwide index. The variances are summed, accounting for the square of the weights, and converted to log(SE) for use in the assessment model assuming log-normal error. There were geographically limited surveys conducting during 1963-1989, with summarized catch rates, but no variance estimates available from 1977 (Table 1). For the period prior to 1993 where there are no variance estimates, twice the recent average value is used, and for the coastwide series where spatial coverage is incomplete values are doubled again.

Commercial fishery CPUE (generally referred to as Weight-Per-Unit-Effort or WPUE as landings are recorded in weight) is reported through mandatory logbooks (voluntary only for vessels under 26 feet, 7.9 m, in length), collected by IPHC port samplers, or returned directly to the IPHC by mail. Commercial CPUE is available as far back as the early 1900s (Stewart and Webster 2022) providing a valuable historical record, but spanning a period of continuous fishery development and change, including an important transition to circle hooks in 1984 that substantially increased average catchability (Table 2-4).

Table 1. Modelled survey Numbers-Per-Unit-Effort (NPUE) and log(SE) 1993-2021, raw average observed NPUE 1977-1986; assumed values in italics.

	Region 2		Region 3		Region 4		Region 4B		Coastwide	
Year	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)
1977	0.60	*0.109*	2.00	*0.108*	--	--	--	--	1.47	*0.153*
1978	0.80	*0.109*	1.30	*0.108*	--	--	--	--	1.11	*0.153*
1979	--	--	1.90	*0.108*	--	--	--	--	--	--
1980	1.20	*0.109*	2.50	*0.108*	--	--	--	--	2.01	*0.153*
1981	0.80	*0.109*	3.80	*0.108*	--	--	--	--	2.67	*0.153*
1982	1.84	*0.109*	3.80	*0.108*	--	--	--	--	2.87	*0.153*
1983	2.30	*0.109*	3.40	*0.108*	--	--	--	--	2.88	*0.153*
1984	6.74	*0.109*	11.60	*0.108*	--	--	--	--	9.30	*0.153*
1985	5.65	*0.109*	11.90	*0.108*	--	--	--	--	8.94	*0.153*
1986	4.54	*0.109*	7.80	*0.108*	--	--	--	--	6.26	*0.153*
1993	6.19	0.102	24.17	0.123	1.90	0.147	10.34	0.329	7.40	0.088
1994	7.42	0.106	23.80	0.100	2.16	0.127	10.58	0.298	7.70	0.070
1995	8.85	0.074	25.55	0.089	2.15	0.119	10.80	0.244	8.28	0.061
1996	7.90	0.059	26.35	0.059	2.34	0.099	11.05	0.187	8.41	0.043
1997	7.19	0.055	28.31	0.055	2.54	0.069	11.23	0.115	8.81	0.039
1998	6.13	0.055	24.55	0.056	2.65	0.069	11.17	0.114	7.96	0.039
1999	5.09	0.053	23.82	0.058	2.38	0.073	9.47	0.125	7.41	0.041
2000	5.61	0.054	25.66	0.050	2.50	0.069	8.64	0.132	7.88	0.037
2001	6.49	0.052	22.46	0.050	2.35	0.066	6.74	0.161	7.20	0.036
2002	6.45	0.050	24.98	0.046	2.26	0.069	4.92	0.178	7.56	0.034
2003	5.52	0.052	24.64	0.050	2.16	0.069	4.08	0.206	7.24	0.037
2004	5.06	0.053	27.74	0.048	2.15	0.068	3.83	0.201	7.76	0.037
2005	5.53	0.053	23.25	0.048	2.23	0.068	3.68	0.208	6.99	0.036
2006	5.47	0.051	22.29	0.049	2.31	0.061	4.25	0.192	6.87	0.035
2007	6.09	0.053	23.75	0.048	2.26	0.064	5.42	0.178	7.28	0.035
2008	6.08	0.051	21.49	0.049	2.51	0.069	5.22	0.176	6.97	0.034
2009	6.17	0.052	20.14	0.049	2.49	0.065	4.40	0.188	6.67	0.034
2010	6.16	0.051	20.48	0.048	2.39	0.062	4.17	0.188	6.66	0.034
2011	6.16	0.049	20.78	0.048	2.27	0.061	4.21	0.173	6.65	0.034
2012	7.20	0.048	21.20	0.046	2.22	0.057	3.84	0.184	6.85	0.031
2013	6.97	0.047	16.45	0.046	2.01	0.058	5.29	0.146	5.82	0.031
2014	7.21	0.046	19.31	0.044	2.04	0.051	4.72	0.163	6.42	0.030
2015	7.96	0.048	19.43	0.044	2.07	0.054	4.69	0.149	6.57	0.030
2016	8.10	0.046	19.80	0.046	1.96	0.056	5.25	0.137	6.63	0.031
2017	5.85	0.045	13.99	0.042	1.82	0.061	4.11	0.090	4.98	0.028
2018	5.19	0.043	12.75	0.042	1.71	0.063	4.30	0.137	4.58	0.029
2019	5.30	0.045	11.53	0.044	1.70	0.066	4.31	0.166	4.34	0.031
2020	4.98	0.046	11.85	0.046	1.65	0.083	4.34	0.204	4.33	0.034
2021	5.72	0.046	15.19	0.048	1.60	0.071	4.25	0.183	5.08	0.034

Table 2. Commercial fishery Weight-Per-Unit-Effort (WPUE) 1907-1949 and estimated log(SE); assumed values in italics.

	Region 2		Region 3		Region 4		Region 4B		Coastwide	
Year	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)
1907	280.00	*0.100*	--	--	--	--	--	--	280.00	*0.100*
1910	271.00	*0.100*	--	--	--	--	--	--	271.00	*0.100*
1911	237.00	*0.100*	--	--	--	--	--	--	237.00	*0.100*
1912	176.00	*0.100*	--	--	--	--	--	--	176.00	*0.100*
1913	128.94	*0.100*	--	--	--	--	--	--	129.00	*0.100*
1914	124.13	*0.100*	--	--	--	--	--	--	124.00	*0.100*
1915	118.02	*0.100*	266.10	*0.100*	--	--	--	--	118.00	*0.100*
1916	114.60	*0.100*	202.80	*0.100*	--	--	--	--	137.00	*0.100*
1917	81.80	*0.100*	157.90	*0.100*	--	--	--	--	98.00	*0.100*
1918	87.50	*0.100*	125.40	*0.100*	--	--	--	--	96.00	*0.100*
1919	82.30	*0.100*	129.90	*0.100*	--	--	--	--	93.00	*0.100*
1920	84.10	*0.100*	147.90	*0.100*	--	--	--	--	96.00	*0.100*
1921	76.46	*0.100*	141.17	*0.100*	--	--	--	--	88.00	*0.100*
1922	62.44	*0.100*	133.79	*0.100*	--	--	--	--	73.00	*0.100*
1923	56.68	*0.100*	149.97	*0.100*	--	--	--	--	78.00	*0.100*
1924	55.39	*0.100*	109.13	*0.100*	--	--	--	--	74.00	*0.100*
1925	51.21	*0.100*	94.63	*0.100*	--	--	--	--	68.00	*0.100*
1926	51.67	*0.100*	93.73	*0.100*	--	--	--	--	67.00	*0.100*
1927	48.83	*0.100*	86.32	*0.100*	--	--	--	--	65.00	*0.100*
1928	47.27	*0.100*	72.34	*0.100*	--	--	--	--	58.00	*0.100*
1929	38.55	*0.100*	70.79	*0.100*	--	--	--	--	51.00	*0.100*
1930	34.44	*0.100*	65.91	*0.100*	--	--	--	--	46.00	*0.100*
1931	38.48	*0.100*	76.17	*0.100*	--	--	--	--	50.00	*0.100*
1932	47.50	*0.100*	83.49	*0.100*	--	--	--	--	60.00	*0.100*
1933	50.16	*0.100*	83.99	*0.100*	--	--	--	--	63.00	*0.100*
1934	54.07	*0.100*	74.97	*0.100*	--	--	--	--	62.00	*0.100*
1935	61.77	*0.100*	97.57	*0.100*	--	--	--	--	76.00	*0.100*
1936	54.66	*0.100*	96.70	*0.100*	--	--	--	--	71.00	*0.100*
1937	61.48	*0.100*	109.99	*0.100*	--	--	--	--	80.00	*0.100*
1938	70.33	*0.100*	114.29	*0.100*	--	--	--	--	88.00	*0.100*
1939	61.90	*0.100*	112.21	*0.100*	--	--	--	--	80.00	*0.100*
1940	61.71	*0.100*	116.38	*0.100*	--	--	--	--	81.00	*0.100*
1941	62.54	*0.100*	122.26	*0.100*	--	--	--	--	85.00	*0.100*
1942	65.43	*0.100*	132.54	*0.100*	--	--	--	--	90.00	*0.100*
1943	72.24	*0.100*	131.27	*0.100*	--	--	--	--	95.00	*0.100*
1944	86.84	*0.100*	149.23	*0.100*	--	--	--	--	110.00	*0.100*
1945	79.69	*0.100*	130.86	*0.100*	--	--	--	--	102.00	*0.100*
1946	83.78	*0.100*	123.82	*0.100*	--	--	--	--	101.00	*0.100*
1947	86.30	*0.100*	114.56	*0.100*	--	--	--	--	99.00	*0.100*
1948	88.61	*0.100*	112.20	*0.100*	--	--	--	--	99.00	*0.100*
1949	85.01	*0.100*	105.89	*0.100*	--	--	--	--	95.00	*0.100*

Table 3. Commercial fishery Weight-Per-Unit-Effort (WPUE) 1950-1991 and estimated log(SE); assumed values in italics.

	Region 2		Region 3		Region 4		Region 4B		Coastwide	
Year	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)
1950	87.66	*0.100*	103.60	*0.100*	--	--	--	--	95.00	*0.100*
1951	87.63	*0.100*	108.93	*0.100*	--	--	--	--	96.00	*0.100*
1952	95.58	*0.100*	128.86	*0.100*	--	--	--	--	110.00	*0.100*
1953	128.65	*0.100*	134.32	*0.100*	--	--	--	--	131.00	*0.100*
1954	137.97	*0.100*	127.43	*0.100*	--	--	--	--	133.00	*0.100*
1955	122.20	*0.100*	116.32	*0.100*	--	--	--	--	119.00	*0.100*
1956	132.02	*0.100*	126.05	*0.100*	--	--	--	--	129.00	*0.100*
1957	100.95	*0.100*	119.84	*0.100*	--	--	--	--	110.00	*0.100*
1958	101.96	*0.100*	139.96	*0.100*	--	--	--	--	121.00	*0.100*
1959	98.67	*0.100*	160.62	*0.100*	--	--	--	--	129.00	*0.100*
1960	105.02	*0.100*	156.08	*0.100*	--	--	--	--	132.00	*0.100*
1961	96.00	*0.100*	159.79	*0.100*	--	--	--	--	127.00	*0.100*
1962	84.76	*0.100*	136.89	*0.100*	--	--	--	--	115.00	*0.100*
1963	77.73	*0.100*	123.89	*0.100*	--	--	--	--	105.00	*0.100*
1964	75.27	*0.100*	120.10	*0.100*	--	--	--	--	100.00	*0.100*
1965	86.47	*0.100*	107.07	*0.100*	--	--	--	--	99.00	*0.100*
1966	82.59	*0.100*	112.72	*0.100*	--	--	--	--	100.00	*0.100*
1967	81.44	*0.100*	113.00	*0.100*	--	--	--	--	101.00	*0.100*
1968	86.58	*0.100*	111.62	*0.100*	--	--	--	--	103.00	*0.100*
1969	81.53	*0.100*	105.07	*0.100*	--	--	--	--	95.00	*0.100*
1970	73.62	*0.100*	103.67	*0.100*	--	--	--	--	91.00	*0.100*
1971	76.05	*0.100*	96.31	*0.100*	--	--	--	--	89.00	*0.100*
1972	69.47	*0.100*	82.87	*0.100*	--	--	--	--	78.00	*0.100*
1973	64.41	*0.100*	62.13	*0.100*	--	--	--	--	63.00	*0.100*
1974	60.89	*0.100*	61.95	*0.100*	--	--	--	--	61.00	*0.100*
1975	61.87	*0.100*	66.76	*0.100*	--	--	--	--	61.00	*0.100*
1976	44.39	*0.100*	61.91	*0.100*	--	--	--	--	55.00	*0.100*
1977	64.17	*0.100*	65.57	*0.100*	--	--	--	--	63.00	*0.100*
1978	54.06	*0.100*	68.47	*0.100*	--	--	--	--	71.00	*0.100*
1979	55.80	*0.100*	67.33	*0.100*	--	--	--	--	75.00	*0.100*
1980	59.54	*0.100*	116.09	*0.100*	--	--	--	--	94.00	*0.100*
1981	73.84	*0.100*	148.86	*0.100*	136.84	*0.100*	99.00	0.078	111.00	*0.100*
1982	71.85	*0.100*	181.34	*0.100*	98.68	*0.100*	--	--	127.00	*0.100*
1984	151.95	0.045	491.33	0.046	386.90	*0.100*	161.00	0.103	316.00	0.035
1985	161.59	0.051	535.06	0.039	456.18	0.099	234.00	0.160	352.00	0.034
1986	137.26	0.035	506.00	0.042	308.70	0.062	238.00	0.372	315.00	0.041
1987	135.53	0.027	490.38	0.036	360.93	0.159	220.00	0.111	316.00	0.038
1988	168.40	0.054	560.55	0.042	405.68	0.105	224.00	0.122	363.00	0.036
1989	154.92	0.042	507.69	0.031	387.41	0.078	268.00	0.094	353.00	0.025
1990	194.64	0.043	403.54	0.036	370.26	0.095	209.00	0.103	315.00	0.029
1991	170.62	0.039	375.02	0.041	367.06	0.157	329.00	0.085	314.00	0.038

Table 4. Commercial fishery Weight-Per-Unit-Effort (WPUE) 1992-2021 and estimated log(SE).

	Region 2		Region 3		Region 4		Region 4B		Coastwide	
Year	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)	Index	log(SE)
1992	167.66	0.040	413.39	0.048	324.01	0.117	280.00	0.095	315.00	0.035
1993	200.04	0.031	439.11	0.096	399.87	0.448	218.00	0.220	369.00	0.100
1994	175.74	0.027	362.77	0.049	343.14	0.333	197.00	0.101	302.00	0.069
1995	190.73	0.025	439.48	0.043	330.22	0.100	189.00	0.336	326.00	0.037
1996	208.81	0.042	505.01	0.046	427.58	0.138	269.00	0.185	387.00	0.039
1997	237.52	0.035	498.02	0.026	417.44	0.107	275.00	0.064	400.00	0.025
1998	221.23	0.029	512.59	0.036	411.86	0.089	287.00	0.058	402.00	0.025
1999	249.48	0.079	475.49	0.024	385.64	0.061	310.00	0.045	390.00	0.023
2000	227.94	0.036	492.21	0.025	403.74	0.082	318.00	0.046	396.00	0.020
2001	202.84	0.039	454.52	0.029	363.00	0.213	270.00	0.076	358.00	0.042
2002	214.81	0.032	466.46	0.025	296.56	0.082	245.00	0.081	356.00	0.020
2003	208.95	0.018	439.27	0.024	251.12	0.072	196.00	0.068	325.00	0.018
2004	192.88	0.028	425.79	0.026	235.23	0.072	202.00	0.061	315.00	0.019
2005	178.98	0.024	387.69	0.023	219.59	0.063	238.00	0.093	293.00	0.017
2006	180.22	0.024	360.70	0.022	178.26	0.064	218.00	0.111	268.00	0.019
2007	158.14	0.023	344.27	0.026	154.65	0.055	230.00	0.108	249.00	0.020
2008	138.83	0.020	318.17	0.024	162.55	0.071	193.00	0.069	229.00	0.017
2009	152.95	0.020	277.22	0.020	174.43	0.055	189.00	0.097	220.00	0.018
2010	185.21	0.037	242.32	0.024	143.97	0.080	142.00	0.063	202.00	0.020
2011	179.95	0.019	226.65	0.025	143.25	0.056	165.00	0.103	196.00	0.015
2012	193.96	0.020	213.46	0.032	139.17	0.080	149.00	0.066	193.00	0.021
2013	192.78	0.026	189.98	0.033	122.70	0.072	127.00	0.064	178.00	0.017
2014	210.44	0.026	182.93	0.039	116.04	0.092	146.00	0.070	183.00	0.022
2015	217.37	0.024	224.46	0.045	136.04	0.065	149.00	0.076	202.00	0.025
2016	212.66	0.019	216.22	0.044	128.30	0.066	123.00	0.083	196.00	0.020
2017	213.02	0.020	219.60	0.037	126.95	0.079	120.00	0.082	202.00	0.020
2018	197.07	0.026	191.12	0.056	115.12	0.058	134.00	0.071	178.00	0.028
2019	186.60	0.030	213.51	0.038	101.85	0.100	115.00	0.084	180.00	0.022
2020	175.93	0.025	216.61	0.041	100.27	0.084	105.00	0.059	178.00	0.022
2021	197.63	*0.055*	206.85	*0.090*	120.82	*0.164*	94.00	*0.152*	182.00	*0.049*

Age data

At each FISS station, otoliths are sampled randomly at rates selected to generate 1500 per IPHC Regulatory Area per year. The number of stations contributing to the annual age information varies considerably over the time-series, with Biological Region 3 the most heavily sampled, followed by Region 2, Region 4 and far fewer samples collected in Region 4B (Table 5). There are also a small number of geographically limited surveys from the period 1963-1966 for which there are age samples, but no corresponding index. Otoliths from the commercial fishery landings are also sampled in proportion to the weight of the catch with different rates by IPHC Regulatory Area (Kong et al. 2022). This has led to a relatively larger number of commercial trips sampled in Biological Region 2 over most of the historical period, with Region 3, Region 4, and Region 4B each contributing fewer samples (Table 6-7).

Table 5. Number of stations contributing to FISS age data (1963-2021).

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
1963	--	236	--	--	236
1964	--	305	--	--	305
1965	121	146	--	--	267
1966	66	--	--	--	66
1977	58	100	--	--	158
1978	62	98	--	--	160
1979	--	104	--	--	104
1980	80	101	--	--	181
1981	72	102	--	--	174
1982	154	148	--	--	302
1983	192	101	--	--	293
1984	241	198	--	--	439
1985	166	103	--	--	269
1986	178	97	--	--	275
1988	72	--	--	--	72
1989	--	33	--	--	33
1993	66	70	--	--	136
1994	14	147	--	--	161
1995	103	120	--	--	223
1996	198	424	--	--	622
1997	211	424	220	74	929
1998	228	507	100	42	877
1999	332	554	61	82	1029
2000	239	548	149	83	1019
2001	330	520	146	83	1079
2002	313	555	154	82	1104
2003	323	516	153	82	1074
2004	327	523	145	70	1065
2005	340	507	144	81	1072
2006	317	526	240	84	1167
2007	330	538	176	73	1117
2008	338	549	166	76	1129
2009	333	537	171	84	1125
2010	333	521	172	76	1102
2011	358	549	166	79	1152
2012	354	522	168	71	1115
2013	364	528	167	78	1137
2014	381	556	227	76	1240
2015	352	529	239	81	1201
2016	350	538	220	72	1180
2017	371	521	166	118	1176
2018	466	537	167	77	1247
2019	482	560	167	81	1290
2020	370	494	--	--	864
2021	393	550	77	37	1057

Table 6. Number of commercial fishing trips contributing to fishery age data (1935-1982); historical values in italics are assumed.

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
1935	*50*	*50*	--	--	*100*
1936	*50*	*50*	--	--	*100*
1937	*50*	*50*	--	--	*100*
1938	*50*	*50*	--	--	*100*
1939	*50*	*50*	--	--	*100*
1940	*50*	*50*	--	--	*100*
1941	*50*	*50*	--	--	*100*
1942	*50*	*50*	--	--	*100*
1943	*50*	*50*	--	--	*100*
1944	*50*	*50*	--	--	*100*
1945	*50*	*50*	*5*	--	*100*
1946	*50*	*50*	*5*	--	*100*
1947	*50*	*50*	*5*	--	*100*
1948	*50*	*50*	*5*	--	*100*
1949	*50*	*50*	*5*	--	*100*
1950	*50*	*50*	*5*	--	*100*
1951	*50*	*50*	*5*	--	*100*
1952	*50*	*50*	*5*	--	*100*
1953	*50*	*50*	*5*	--	*100*
1954	*50*	*50*	*5*	--	*100*
1955	*50*	*50*	*5*	--	*100*
1956	*50*	*50*	*5*	--	*100*
1957	*50*	*50*	*5*	--	*100*
1958	*50*	*50*	*5*	--	*100*
1959	*50*	*50*	*5*	--	*100*
1960	*50*	*50*	*5*	--	*100*
1961	*50*	*50*	*5*	--	*100*
1962	*50*	*50*	*5*	--	*100*
1963	*50*	*50*	*5*	--	*100*
1964	116	100	14	--	230
1965	118	106	12	--	238
1966	102	113	12	--	228
1967	125	133	20	--	278
1968	135	132	14	--	282
1969	113	102	12	--	227
1970	97	125	18	--	241
1971	82	77	9	--	168
1972	552	196	3	--	752
1973	311	262	5	--	578
1974	153	68	3	--	226
1975	234	76	7	--	320
1976	332	135	7	--	476
1977	247	138	7	--	401
1978	241	120	4	--	377
1979	125	101	6	--	244
1980	140	113	1	--	262
1981	146	90	7	--	248
1982	168	137	11	--	316

Table 7. Number of commercial fishing trips contributing to fishery age data (1983-2021).

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
1983	133	106	23	6	268
1984	170	90	9	13	282
1985	171	99	14	2	286
1986	158	152	34	1	345
1987	531	498	76	12	1117
1988	278	258	19	16	571
1989	318	371	39	24	752
1990	491	560	50	3	1104
1991	718	496	62	12	1288
1992	1027	478	61	20	1586
1993	959	471	65	11	1506
1994	896	474	89	31	1490
1995	887	468	72	37	1464
1996	859	437	76	27	1399
1997	676	429	183	58	1346
1998	515	277	127	47	966
1999	454	303	118	24	899
2000	512	358	119	27	1016
2001	505	233	117	13	868
2002	561	284	163	53	1061
2003	545	266	118	49	978
2004	491	200	75	9	775
2005	461	193	125	13	792
2006	483	256	81	22	842
2007	429	218	95	12	754
2008	385	221	98	11	715
2009	432	240	68	14	754
2010	354	260	97	25	736
2011	383	224	83	14	704
2012	421	217	81	13	732
2013	455	196	73	14	738
2014	426	221	64	8	719
2015	476	192	119	15	802
2016	466	164	112	15	757
2017	410	175	106	17	708
2018	337	178	105	17	637
2019	409	199	116	10	734
2020	406	176	47	12	641
2021	272	126	30	7	435

As has been the case since the 2015 stock assessment (Stewart and Martell 2016), all age data used in the stock assessment is aggregated into bins of ages from age-2 to age-25, with age 2 representing a ‘minus’ group including all fish of age 2 and younger, and age 25 representing a ‘plus’ group including all fish age 25 and older. For years prior to 2002 (except the survey ages from 1998 which were re-aged in 2013), surface ages were the standard method, replaced by break-and-bake in recent years. Because surface ages are known to be biased at older ages (Forsberg and Stewart 2015), the age data are aggregated at a lower ‘plus’ group, age 20+, for all years where this was the primary method.

Beginning with the 2019 stock assessment, sex-specific fishery age data has been available via the collection of fin clips and subsequent genetic assay based on sampling begun in 2017. The processing of these samples lags one-year, thus for the 2021 stock assessment there were four years of sex-specific fishery age compositions used (2017-2020). They are compiled in an identical manner to the standard fishery age data, but delineating males and females through the weighting and aggregation up to Biological Regions and coastwide.

Other biological and fishery information

There are several other sources of information contributing to the stock assessment models. These include:

1) the time-series of the Pacific Decadal Oscillation (PDO) index
2) the maturity ogive
3) fecundity information
4) weight-at-age
5) length-weight relationship
6) ageing error (bias and imprecision)
7) data based ‘priors’ on bycatch, discard, and recreational selectivity

The only significant changes to the treatment of these sources of information since the 2015 stock assessment (Stewart and Martell 2016), is the introduction of a revised length-weight relationship in 2021 (Webster and Stewart 2022). Because the directly measured weights collected during the FISS (since 2019) and the commercial sampling (2015) have been used directly in the stock assessment data preparation, the updated length-weight relationship has little effect on the assessment, except through potentially more accurate calculations by domestic agencies of mortality in weight from piece counts (this is relevant to non-directed discard mortality, recreational mortality and subsistence mortality). These effects will be realized gradually as calculation routines are updated and data sources are reported to the IPHC.

All other sources of information are updated (where appropriate) and described each year in the annual overview of data sources (Stewart and Webster 2022). For convenience, the treatment of each is briefly summarized in Table 8.

Table 8. Summary of other information sources contributing directly to stock assessment input files (Stewart and Webster 2022).

Input	Summary	Key assumptions
Pacific Decadal Oscillation index[1]	Monthly values averaged and compiled into a binary index for each year based on assignment to 'positive' and 'negative' phases	Used as a binary indicator rather than annually varying values (but see sensitivity analyses below).
Maturity	Trimmed logistic from Clark and Hare (2006); 50% female maturity at 11.6 years old.	Based on visual assessments, treated as age-based and time-invariant.
Fecundity	Assumed to be proportional to body weight.	Temporal variability is included via changes in weight-at-age.
Weight-at-age	Reconstructed from survey and fishery information by Biological Region.	Historical variability has been similar for female and male Pacific halibut.
Length-weight relationship	Not used directly in the assessment, most of the historical data relies on a constant average length-weight relationship.	Measured weights are used preferentially where available.
Ageing error	Pacific halibut are relatively easy to age accurately and with a high degree of precision using the break-and-bake method (Clark 2004a, 2004b; Clark and Hare 2006; Piner and Wischnioski 2004). Surface ages are biased and less precise (Stewart 2014).	Multi-decadal comparison suggest that accuracy and precision have not changed appreciably over the entire historical record (Forsberg and Stewart 2015).
Bycatch selectivity prior	Age-distributions are created from weighted and aggregated length frequencies from a variety of sources and age-length keys from trawl surveys.	Due to incomplete sampling, poor data quality in many years, and other uncertainties, data are considered unreliable for estimation of recruitment.
Discard selectivity prior	Age-distributions of sub-legal (<32 inch) Pacific Halibut captured by the FISS are used as a proxy for poorly sampled directed commercial fishery discards.	Survey data may not be representative of commercial fishing behavior but are currently the only source of information on the age range of discarded fish.
Recreational selectivity prior	Weighted age-frequency data from the IPHC Regulatory Area 3A recreational fishery are the only comprehensive source available.	These data may not be representative of all recreational mortality but provide the best information currently available.

External information on M

In 2021 a CAPAM workshop on natural mortality (formal report still pending) was held with the objective of developing best practices for the treatment of *M* in stock assessment modelling. Two primary conclusions were evident from the discussions at the workshop:

1) Although results are varied, simulations have generally indicated that estimation of *M* is preferable to the use of fixed values, where this is possible. The use of informative priors is frequently necessary, with the most common prior based on longevity.
2) Elevated *M* at the youngest ages/smallest sizes should be expected due to increased size-dependent predation mortality.

---

[1] Data can be accessed at: https://oceanview.pfeg.noaa.gov/erddap/tabledap/cciea_OC_PDO.htmlTable?time,PDO

For the 2022 Pacific halibut assessment, both of these conclusions were evaluated for inclusion into the four stock assessment models. First, an age-independent prior on *M* for Pacific halibut was developed based on the meta-analysis of Hamel (2014; and subsequently updated, Hamel pers. comm.), which uses the prediction interval based on a meta-analysis of the maximum observed age for a wide range of species. Both male and female Pacific halibut have been observed to age-55 (with multiple fish of both sexes exceeding age-50 indicating that this is likely to be an accurate estimate of longevity, and not an artifact of a single case of ageing imprecision). The prior median is given by:

$$M = \frac{5.4}{Age_{max}}$$

which results in a value of 0.0982, and a log(SD) of 0.438. With such a large variance, this prior is only weakly informative (Figure 4), but still may provide additional stability for estimation of *M*.

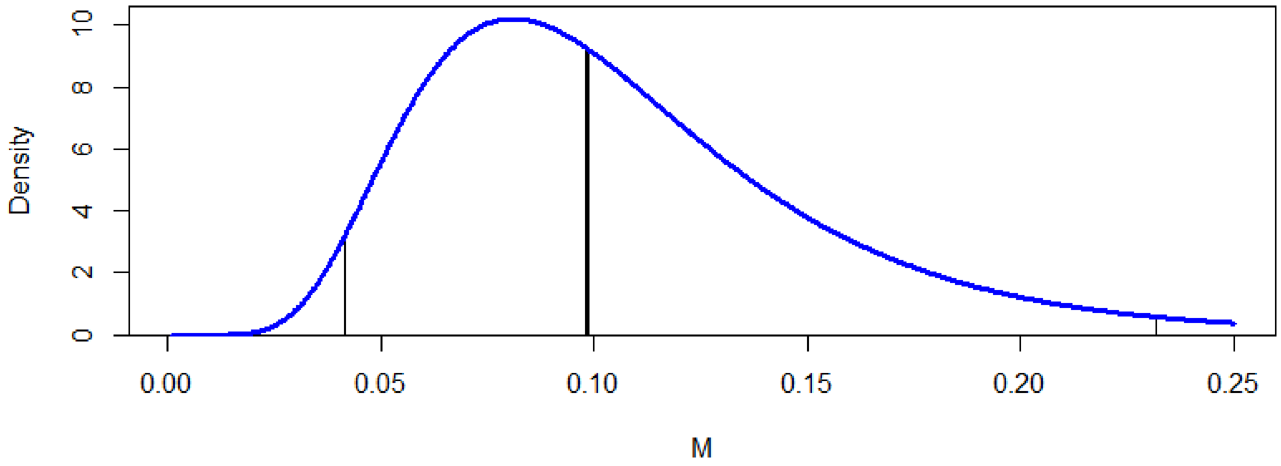


Figure 4. Informative prior for *M*. Thick vertical line denotes the median, thin lines the 2.5 and 97.5 percentiles of the distribution.

To explore the potential that *M* for Pacific halibut should be size-dependent, the average size at age was described from trawl survey data, which provides the best source of information on fish that are too small to be reliably captured with commercial or FISS longline gear. Sexual dimorphism is relatively small at the youngest ages, and rapid growth of both males and females proceeds at approximately 10 cm per year for the first 5-6 years of life (Figure 5).

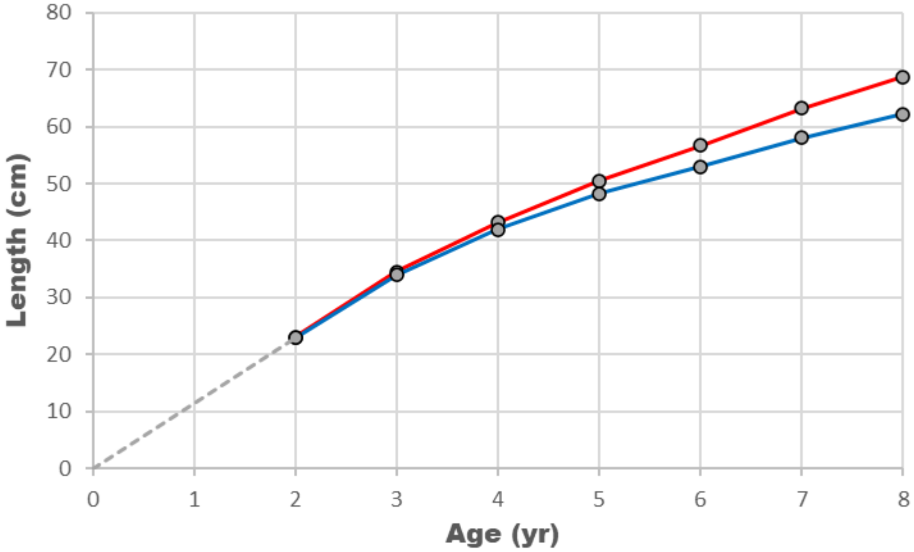


Figure 5. Average Pacific halibut length-at-age based on recent trawl survey data; the red line denotes females, and the blue line denotes males. The dashed line is a simple extrapolation for unobserved ages 0-1, assuming zero length at age-0.

Because of their very rapid growth, it might be expected that Pacific halibut would endure a lower *M* at age than other flatfish congeners, and so any meaningful comparison might be best summarized in terms of size. To explore how Pacific halibut size-at-age and *M* compare to other flatfish species, a summary of all available Northeast Pacific flatfish stock assessments was conducted. For each assessment, the estimate or fixed value of *M* and the average asymptotic size (either L_{inf} or L_{old}, depending on the parameterization) was recorded, separated by males and females where possible, as dimorphic growth is relatively common among flatfish. Complete data were available for 26 stocks spread over four geographical regions comprising the U.S. West Coast, British Columbia, Gulf of Alaska, and the Bering Sea (Table 8, Figure 6). These stocks represented 14 individual species, of which all but 5 reported sex-specific *M* and maximum size. There was no clear pattern of higher *M* for smaller flatfish, although the highest *M* values all occurred for flatfish with asymptotic size of approximately 60 cm (Figure 6) and for every stock with separate *M* values by sex, the higher *M* was associated with a smaller maximum size. A key result of this comparison is that flatfish with maximum sizes of >35 cm had natural mortality values both above and below those used for Pacific halibut. As Pacific halibut reach this size by age-3, this suggests that strongly elevated natural mortality due to predation common across flatfish species is likely to occur primarily below this age. Also of note is that the Hamel prior for Pacific halibut derived above contains the majority of the flatfish species examined within the 95% prediction interval.

Table 9. Summary of M and L_{inf} values from all available Northeast Pacific flatfish stocks. Where sex-specific values were not reported the combined value is included in the female column; in some cases only females were modelled and male values are missing. Region abbreviations: C GOA indicates the central Gulf of Alaska, W GOA the Western Gulf and E Goa the Eastern Gulf; BSAI indicates the Bering Sea and Aleutian Islands, BC the waters off British Columbia and WC the waters off the west coast of the continental United States.

		M		L_{inf}		
Species	Region	female	male	female	male	Reference
Northern rock sole	C GOA	0.200	0.232	50.29	41.92	Bryan and Palsson (2021)
Northern rock sole	W GOA	0.200	0.254	45.47	37.72	
Southern rock sole	C GOA	0.200	0.253	51.43	39.86	
Southern rock sole	W GOA	0.200	0.271	48.67	39.15	
Flathead sole	GOA	0.200	--	44.40	--	Turnock et al. (2017)
Arrowtooth flounder	GOA	0.200	0.350	83.76	52.41	Shotwell et al. (2021)
Rex sole	W GOA	0.170	0.170	46.83	41.02	McGilliard and Palsson (2021)
Rex sole	E GOA	0.170	0.170	36.73	34.64	
Dover sole	GOA	0.113	0.119	50.75	43.44	McGilliard et al. (2019)
Yellowfin sole	GOA	0.200	--	34.00	--	Bryan and Ferriss (2021)
Alaska plaice	BSAI	0.130	0.130	50.10	49.90	Ormseth (2021)
Flathead sole	BSAI	0.200	0.200	44.88	37.57	Monnahan and Haehn (2020)
Kamchatka flounder	BSAI	0.110	0.110	79.60	60.73	Bryan et al. (2020a)
Arrowtooth flounder	BSAI	0.200	0.350	84.83	52.70	Shotwell et al. (2020)
Greenland turbot	BSAI	0.112	0.112	90.29	71.99	Bryan et al. (2020b)
Yellowfin sole	BSAI	0.120	0.135	38.03	34.03	Spies et al. (2021)
Rock sole	BC	0.200	--	50.50	--	Holt et al. (2016)
English sole	BC	0.200	--	49.40	--	Starr (2009b)
Petrale sole	BC	0.210	0.210	56.30	45.80	Starr (2009a)
Arrowtooth flounder	BC	0.328	--	60.90	47.80	Grandin and Forrest (2017)
Dover sole	WC	0.108	0.114	48.05	41.98	Wetzel and Berger (2021)
Petrale sole	WC	0.159	0.164	53.12	40.83	Wetzel (2019)
Arrowtooth flounder	WC	0.216	0.300	69.77	44.40	Sampson et al. (2017)
Pacific sanddab	WC	0.459	0.566	30.33	26.47	He et al. (2013)
English sole	WC	0.260	0.260	40.56	23.99	Stewart (2007)
Starry flounder	WC	0.500	0.750	59.10	49.70	Ralston (2005)

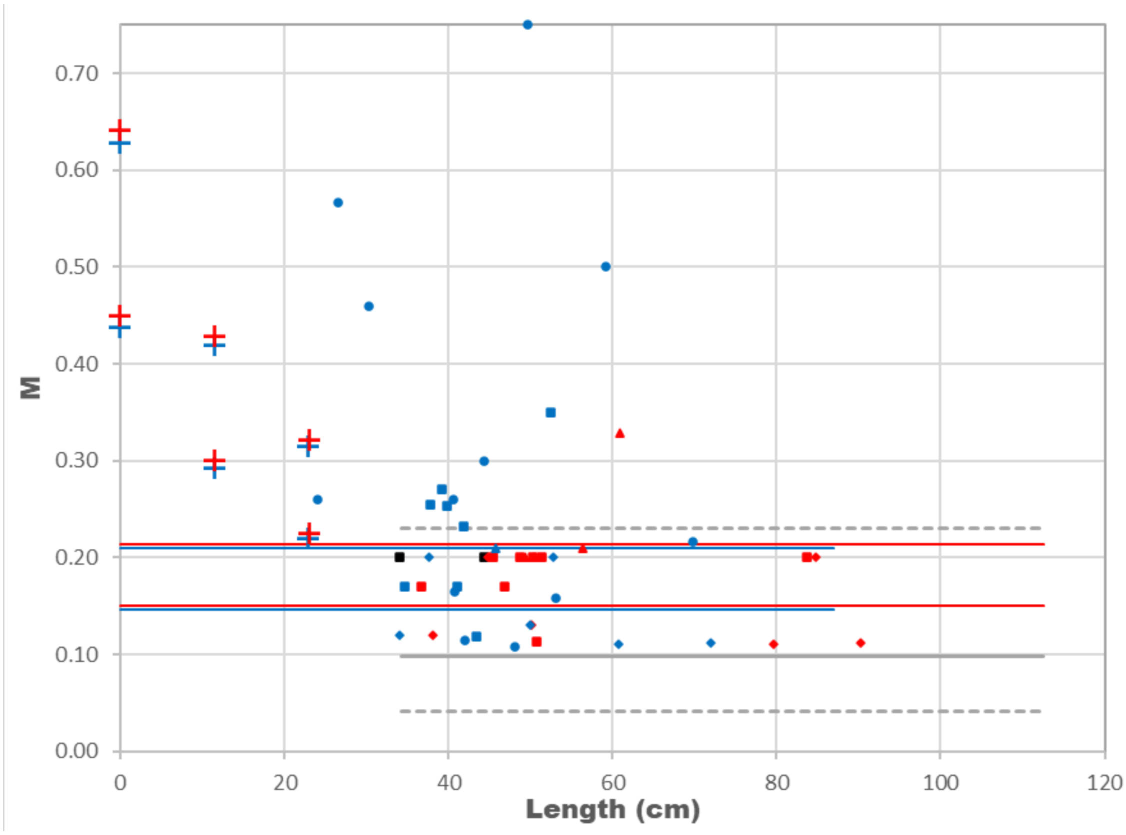


Figure 6. Average natural *M* for 26 Northeast Pacific flatfish stocks: males denoted by blue symbols, females by red symbols, combined sexes by black symbols. Regions are indicated by the point type: diamonds are the Bering Sea, squares are the Gulf of Alaska, triangles are British Columbia, and circles are the U.S. West Coast. The Hamel prior for Pacific halibut is shown as the grey line, with the dashed grey lines representing the 95% prediction interval. Solid red and blue lines denote the highest and lowest 2021 stock assessment estimates of *M* for ages 0-8, females in red, males in blue. Crosses denote the *M* at the average size for Pacific halibut ages 0-2 proposed for use in the 2022 stock assessment.

With very little data to inform a consistent level of *M* for Pacific halibut less than 35 cm (corresponding to ages 0-2) it was necessary to consider other sources of information. Ecosystem models that include predator-prey dynamics generally suggest much higher *M* for the youngest age classes of NE Pacific groundfish (Adams et al. 2022). Where this information has been applied in other assessments used for management advice somewhat arbitrary scalars such as 1.5 x *M* for age 2, 2 x *M* for age-1 and *3* x *M* for age-0 are generally consistent with ecosystem models (e.g., Ianelli et al. 2021). Applying this general approach to Pacific halibut would allow for size-dependent *M* that is consistent with theoretical concepts (Figure 6) but does not appreciably change the natural mortality used for ages represented in observed fishery and survey data (exclusively age 2+). With little to no data at these youngest ages, any effect is likely to ‘scale out’ in the absolute estimates of recruitment deviations; however, when an index of recruitment is evaluated (i.e., the PDO in this assessment; see sensitivity analyses below) it may be important to include elevated *M* at these ages.

Bootstrapping input sample sizes for age compositions

Data weighting in the Pacific halibut stock assessment has historically relied on the number of sampled FISS stations and number of sampled commercial fishery trips as a starting point for all models. Investigation of alternative tuning procedures and likelihoods has been necessarily conditioned on these starting values, yet they had not been evaluated specifically. Following the method developed in Stewart and Hamel (2014) effective sample sizes based on the actual distribution and weighting of both the samples and the fish within samples were bootstrapped for use as inputs to the 2022 stock assessment. Briefly, this method randomly resamples FISS stations (or commercial trips) with replacement from each stratum (IPHC Regulatory Areas for FISS data), then randomly resamples fish within those samples with replacement. Each bootstrapped data set is then used to construct a new age composition. The new age composition is then compared to the actual, and the effective sample size (McAllister and Ianelli 1997; Stewart and Hamel 2014) is calculated. From a set of bootstraps, the harmonic mean of the effective sample size provides an unbiased estimate of the central tendency, provided that sufficient bootstraps have been conducted to avoid appreciable Monte-Carlo error (in this case 10,000 was found to produce <0.5% variability in replicate data sets).

The effective sample size calculated in this manner is analogous to a minimum variance estimate – the actual effective sample size may be lower than calculated if not all strata are fully sampled (measurement error), of the source of the data differs from that assumed in the assessment (structural or process error). However, the effective sample size cannot be larger than the bootstrapped value simply due to the among and within sample variability and the sample sizes achieved. Thus, although time-consuming to produce, the approach provides an objective starting point for data weighting, and a logical upper bound on sample sizes used in the stock assessment models.

For Pacific halibut, the results of this bootstrapping analysis indicated that the effective sample size across all composition data was approximately four times the raw number of samples collected, albeit with considerable interannual variability (Figure 7). Important differences were evident between the fishery data and the FISS data and among geographical aggregations (Table 10). FISS age compositions tended to have slightly lower effective sample sizes per sample than the commercial fishery, consistent with fishery samples representing entire trips, potentially fishing several locations with the fish mixed before sampling occurs at the dock. To the degree that fish school by size and age, it is expected that all fish in each sample will not be independent (e.g., Pennington and Volstad 1994) and thus the effective sample size will tend to be less than the nominal sample size but still increase as additional fish are added to the sample over some range, until the clustering of fish makes additional samples necessary to increase the effective sample size further. In some cases where clustering occurs at a broader scale than the sampling (e.g., young/small fish in one area, old/large fish in another, even samples are not independent and thus the effective sample size can be less than the number of samples. This was observed for Biological Region 4B and was particularly pronounced for the FISS data (Table 10, Figure 8-9).

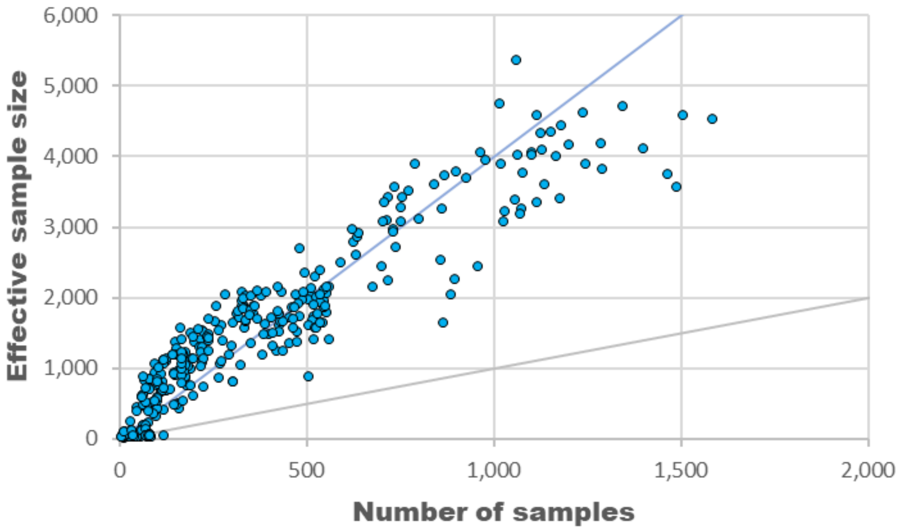


Figure 7. Number of samples vs. bootstrapped effective sample size for all FISS and fishery age compositions data. Grey line indicates a 1:1 relationship, blue line indicates a 4:1 relationship.

Table 10. Summary of bootstrapping results by data type and spatial aggregation.

Data type	Aggregation	Mean effective N per sample
FISS	Coastwide	3.7
FISS	Region 2	5.1
FISS	Region 3	4.1
FISS	Region 4	6.5
FISS	Region 4B	0.6
All fishery	Coastwide	4.1
All fishery	Region 2	3.6
All fishery	Region 3	5.6
All fishery	Region 4	8.8
All fishery	Region 4B	2.6
Sexed fishery	Coastwide	4.1
Sexed fishery	Region 2	3.9
Sexed fishery	Region 3	5.5
Sexed fishery	Region 4	7.9
Sexed fishery	Region 4B	4.4
Unsexed fishery	Coastwide	4.1
Unsexed fishery	Region 2	3.6
Unsexed fishery	Region 3	5.7
Unsexed fishery	Region 4	8.9
Unsexed fishery	Region 4B	2.3

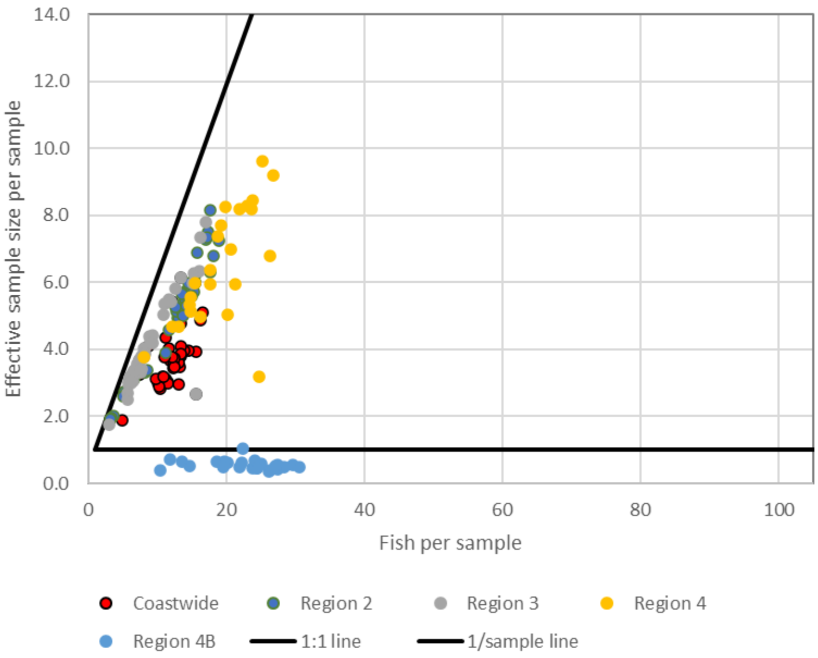


Figure 8. Effective sample size per FISS station sampled for age data as a function of the number of fish sampled by Biological region and coastwide. Diagonal line indicates complete independence among fish within a sample, horizontal line indicates clustering such that fish within samples are not independent. See Stewart and Hamel (2014) for more information.

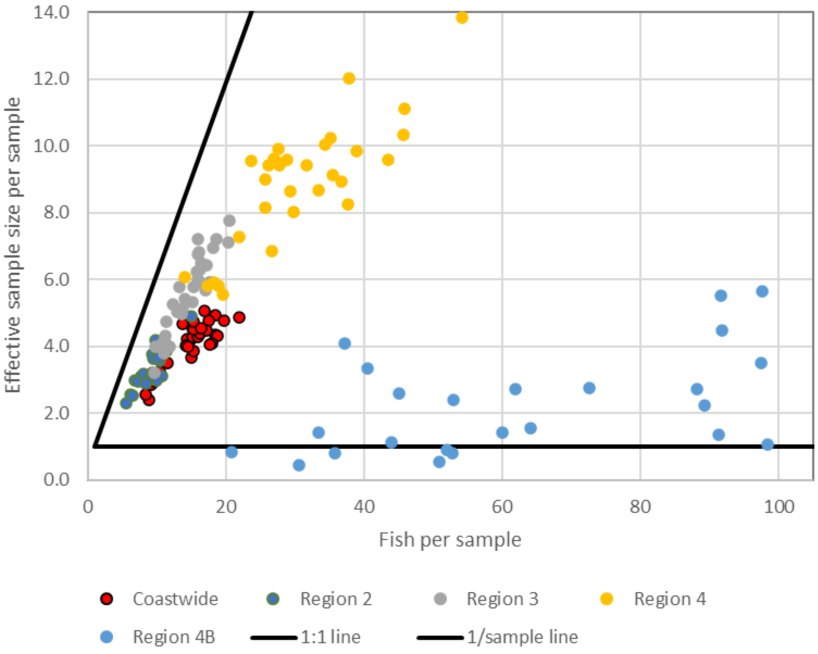


Figure 9. Effective sample size per commercial trip sampled for sexes-aggregated age data as a function of the number of fish sampled by Biological region and coastwide. Diagonal line indicates complete independence among fish within a sample, horizontal line indicates clustering such that fish within samples are not independent. See Stewart and Hamel (2014) for more information.

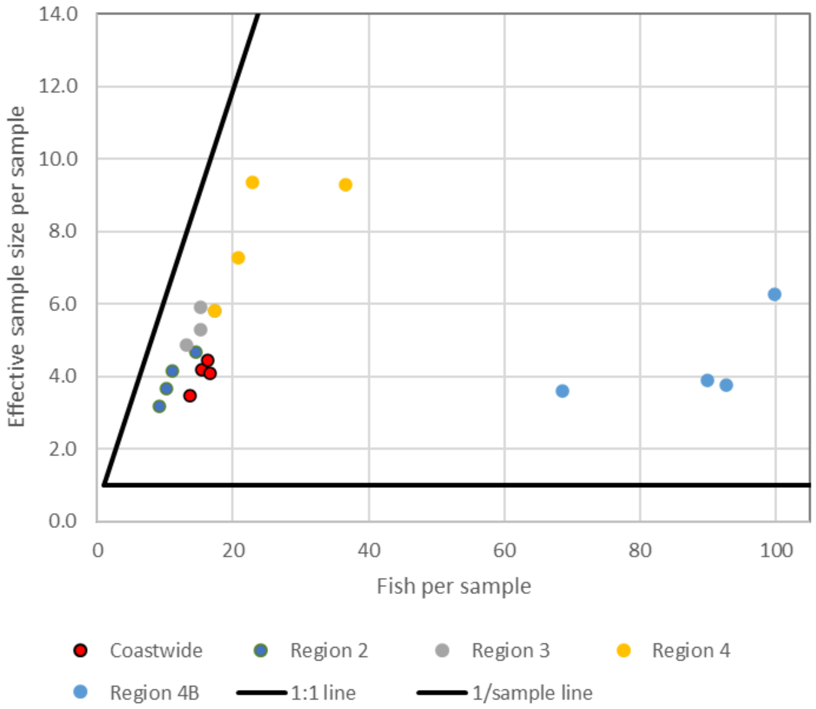


Figure 10. Effective sample size per commercial trip sampled for sex-specific ages as a function of the number of fish sampled by Biological region and coastwide. Diagonal line indicates complete independence among fish within a sample, horizontal line indicates clustering such that fish within samples are not independent. See Stewart and Hamel (2014) for more information.

Because early fishery data are unavailable in current IPHC data bases, age compositions prior to 1991 were unable to be bootstrapped. Instead, the average relationship between the number of samples and the bootstrapped effective sample size (Table 10) was used to approximate effective sample sizes for use as starting values in the assessment models. Bootstrapped FISS (Table 11) and fishery (Table 12-13) effective sample sizes are provided below.

Table 11. Bootstrapped effective sample size for FISS age data (1963-2021).

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
1963	--	1,448	--	--	1,448
1964	--	814	--	--	814
1965	403	479	--	--	866
1966	180	--	--	--	180
1977	117	403	--	--	506
1978	121	309	--	--	433
1979	--	418	--	--	418
1980	216	541	--	--	744
1981	186	797	--	--	946
1982	480	938	--	--	1,313
1983	746	589	--	--	1,181
1984	1,384	599	--	--	1,239
1985	1,127	567	--	--	1,057
1986	1,229	525	--	--	1,091
1988	139	--	--	--	139
1989	--	121	--	--	121
1993	481	514	--	--	692
1994	105	921	--	--	962
1995	839	716	--	--	1,086
1996	1,434	2,141	--	--	2,970
1997	1,548	1,796	1,226	42	3,685
1998	729	882	347	29	1,640
1999	1,903	1,779	586	50	3,225
2000	1,484	1,942	1,370	49	3,888
2001	2,082	1,725	1,194	54	3,769
2002	1,776	2,049	1,275	53	4,056
2003	1,721	1,683	1,180	39	3,253
2004	1,877	2,297	1,189	46	4,025
2005	1,676	1,595	1,187	54	3,190
2006	1,805	2,000	1,233	34	4,002
2007	1,943	2,386	1,047	41	4,575
2008	2,027	1,862	1,401	31	4,084
2009	1,989	1,927	542	31	4,335
2010	1,831	1,886	1,200	37	4,019
2011	1,765	2,107	1,224	37	4,353
2012	1,819	1,568	897	36	3,348
2013	1,868	1,560	782	37	3,605
2014	2,018	2,145	1,126	47	4,620
2015	2,015	1,761	1,431	36	4,170
2016	1,751	2,036	1,030	35	4,429
2017	1,696	1,399	985	47	3,405
2018	1,572	1,637	1,064	36	3,899
2019	2,692	1,403	1,132	41	3,819
2020	2,098	2,070	--	--	3,247
2021	2,068	1,885	387	39	3,382

Table 12. Bootstrapped effective sample size for commercial fishery age data (1964-2011).

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
1964	*412*	*565*	*124*	--	*935*
1965	*419*	*599*	*107*	--	*968*
1966	*362*	*639*	*107*	--	*927*
1967	*444*	*752*	*178*	--	*1,130*
1968	*480*	*746*	*124*	--	*1,146*
1969	*402*	*577*	*107*	--	*923*
1970	*345*	*707*	*160*	--	*980*
1971	*291*	*435*	*80*	--	*683*
1972	*1,961*	*1,108*	*27*	--	*3,057*
1973	*1,105*	*1,481*	*44*	--	*2,350*
1974	*544*	*384*	*27*	--	*919*
1975	*831*	*430*	*62*	--	*1,301*
1976	*1,180*	*763*	*62*	--	*1,935*
1977	*878*	*780*	*62*	--	*1,630*
1978	*856*	*678*	*36*	--	*1,533*
1979	*444*	*571*	*53*	--	*992*
1980	*497*	*639*	*9*	--	*1,065*
1981	*519*	*509*	*62*	--	*1,008*
1982	*597*	*775*	*98*	--	*1,285*
1983	*473*	*599*	*204*	--	*1,090*
1984	*604*	*509*	*80*	--	*1,146*
1985	*608*	*560*	*124*	--	*1,163*
1986	*561*	*859*	*302*	--	*1,403*
1987	*1,887*	*2,816*	*675*	--	*4,541*
1988	*988*	*1,459*	*169*	--	*2,321*
1989	*1,130*	*2,098*	*346*	--	*3,057*
1990	*1,745*	*3,166*	*444*	--	*4,488*
1991	2,242	2,350	593	49	4,181
1992	3,069	1,907	604	48	4,519
1993	2,446	2,031	617	30	4,575
1994	2,258	1,521	516	80	3,560
1995	2,032	1,861	436	31	3,744
1996	2,532	1,660	448	38	4,116
1997	2,148	1,770	1,017	26	4,700
1998	2,035	1,391	1,096	38	4,054
1999	1,713	1,637	1,110	27	3,776
2000	2,133	1,878	1,122	24	4,741
2001	1,967	1,454	802	14	3,727
2002	2,155	2,042	1,569	43	5,359
2003	1,645	1,540	1,061	26	3,944
2004	1,975	1,364	766	30	3,512
2005	1,862	1,497	1,141	29	3,892
2006	1,737	1,661	703	34	3,595
2007	1,607	1,400	1,054	27	3,280
2008	1,486	1,536	876	32	3,086
2009	1,651	1,702	818	19	3,078
2010	1,349	1,875	928	17	3,569
2011	1,190	1,510	833	20	3,072

Table 13. Bootstrapped effective sample size for commercial fishery age data (2012-2021). 2017-2020 represent bootstrapping of the sex-specific age data.

Year	Region 2	Region 3	Region 4	Region 4B	Coastwide
2012	1,248	1,300	837	19	2,970
2013	1,353	1,129	719	38	2,712
2014	1,514	1,297	885	27	3,416
2015	1,373	987	1,120	43	3,105
2016	1,689	966	912	67	3,414
2017	1,316	1,100	1,013	96	3,342
2018	1,655	883	763	47	2,860
2019	1,500	1,127	673	55	2,936
2020	1,713	1,031	388	42	2,917
2021	1,605	672	241	31	2,076

Mortality due to marine mammal depredation

Recent stock assessments have investigated the sensitivity to additional mortality due to marine mammal depredation. Adding mortality generally results in an increase in the scale of the estimated population size: unmodelled whale depredation effectively represents unobserved productivity (Figure 19 in Stewart and Hicks 2022). However, if trending rapidly, this unobserved mortality could also result in bias in the estimated population trend.

The sablefish stock assessment in Alaska (Goethel et al. 2021) accounts for marine mammal depredation by first estimating the effect on catch rates based on the difference in catch rates for depredated compared to non-depredated commercial fishing activity within spatial strata (Peterson et al. 2017; Peterson and Carothers 2013; Peterson et al. 2014). This approach implicitly assumes that depredation is independent of underlying population density and subsequent catch rates. These estimates are then combined with a frequency of interaction estimated from observer data and result in a relatively small positive adjustment to total expected mortality (1.5% for 2020-2021, Goethel et al. 2021).

Analysis of FISS marine mammal interactions indicates that the most important marine mammal depredation for Pacific halibut occurs due to sperm whales in IPHC Regulatory Area 3A and orca whales in IPHC Regulatory Area 4A (Webster 2021). When orca whales are present in IPHC Regulatory Area 4A FISS catch rates were estimated to be reduced to 51% of those when whales were not present, and 84%/86% for orca whales and sperm whales in IPHC Regulatory Area 3A. Because this approach is based on the space-time modelling and informed by the full FISS data set, it implicitly accounts for differences in the underlying biomass distribution. One possible path forward to estimating whale depredation in the Pacific halibut fishery would be to use these estimates of catch-rate reduction along with observations of whale interactions from the commercial fishery in order to estimate additional mortality due to marine mammal depredation associated with commercial fishing.

The IPHC added fields to the commercial fishery logbooks in 2017 for reporting of damage to fishing gear/catch (found to be indicative of marine mammal depredation from FISS observations) as well as the number and species of whales if any were observed. Informal results suggest incomplete participation in completing these fields in the logs –because they are still relatively new, because of the potential sensitivity of marine mammal interactions and because harvesters may not perceive a benefit to accurate reporting of this information. Further, there have been challenges in the collection of these data (e.g., consistent use of reporting codes) as well as processing this information in IPHC databases (e.g., accurate delineation of missing data vs. no marine mammal observations). For these reasons, the summaries provided in this section should be considered highly preliminary and will likely be revised in the future.

Preliminary evaluation of logbooks corresponding to commercial fishing sets targeting Pacific halibut suggest that for most IPHC Regulatory Areas a majority of sets have some information recorded and that completeness may be increasing slightly since the fields were added in 2017 (Figure 11). Using a relatively strict criteria that both some gear damage and at least one marine mammal must have been observed, the reported rate of depredation appears to be around 1% (Figure 12). Further delineating by marine mammal species supports FISS observations of orca activity being most important in IPHC Regulatory Area 4A, sperm whale activity being most important in IPHC Regulatory Area 3A and only a small fraction of interactions with pinnipeds (Figure 13). Raw average WPUE for sets identified as depredated vs. those that were not depredated suggests a similar reduction in catch rates to those estimated for FISS data when orcas were the source of depredation (Figure 14). For sperm whale and pinniped depredation, there was no clear reduction in catch rates; this could be explained by these species depredating in areas with higher catch rates than average, issues with categorizing depredation or other factors. are currently suggestive of trends observed in the FISS but appear to be inadequate. Published observer data on marine mammal interactions with commercial longline fisheries targeting Pacific halibut suggests a slightly higher rate than currently reported in logbooks, but considerable variability among years, either actual or due to relatively low observer coverage rates (supplementary table 3 in Dahlheim et al. 2022).

In aggregate, this preliminary evaluation of depredation suggests that there is some mortality occurring that is not modelled in the current stock assessment, but that it is relatively low, and that the effect is likely to create a slight underestimate of the stock size and productivity. More work is being conducted to determine necessary steps to improve reporting rates, data collection protocols and database issues. Pending these efforts, no formal correction to the mortality time-series is proposed for 2022.

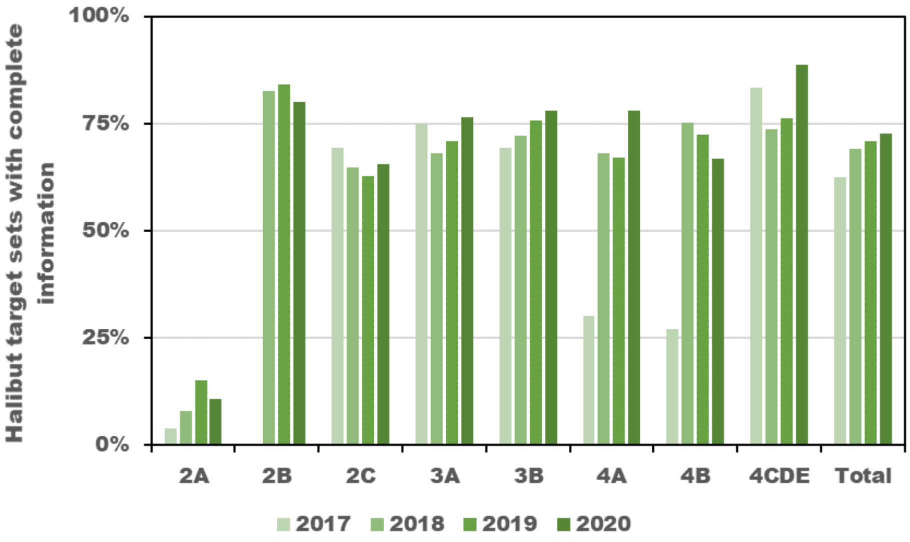


Figure 11. Percent of logbook-recorded sets with apparently complete information by IPHC Regulatory Area and year.

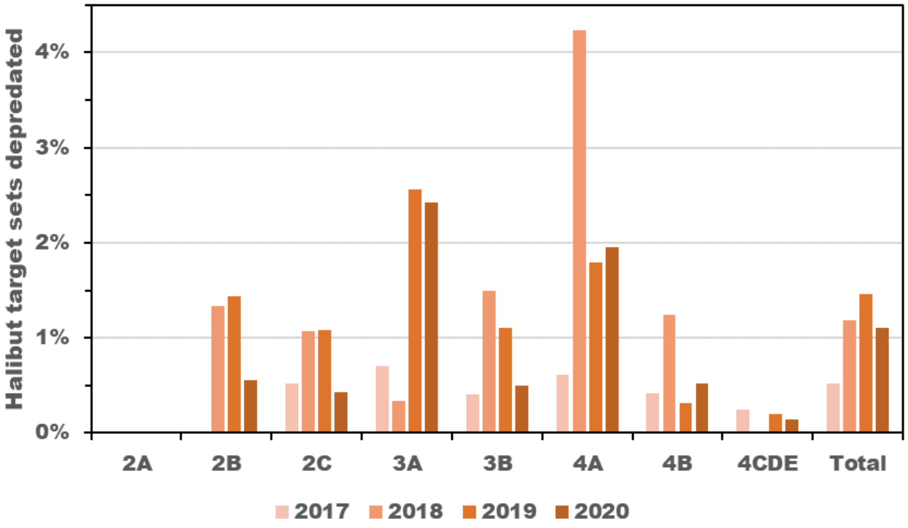


Figure 12. Percent of logbook-recorded sets with reported marine mammal depredation by IPHC Regulatory Area and year.

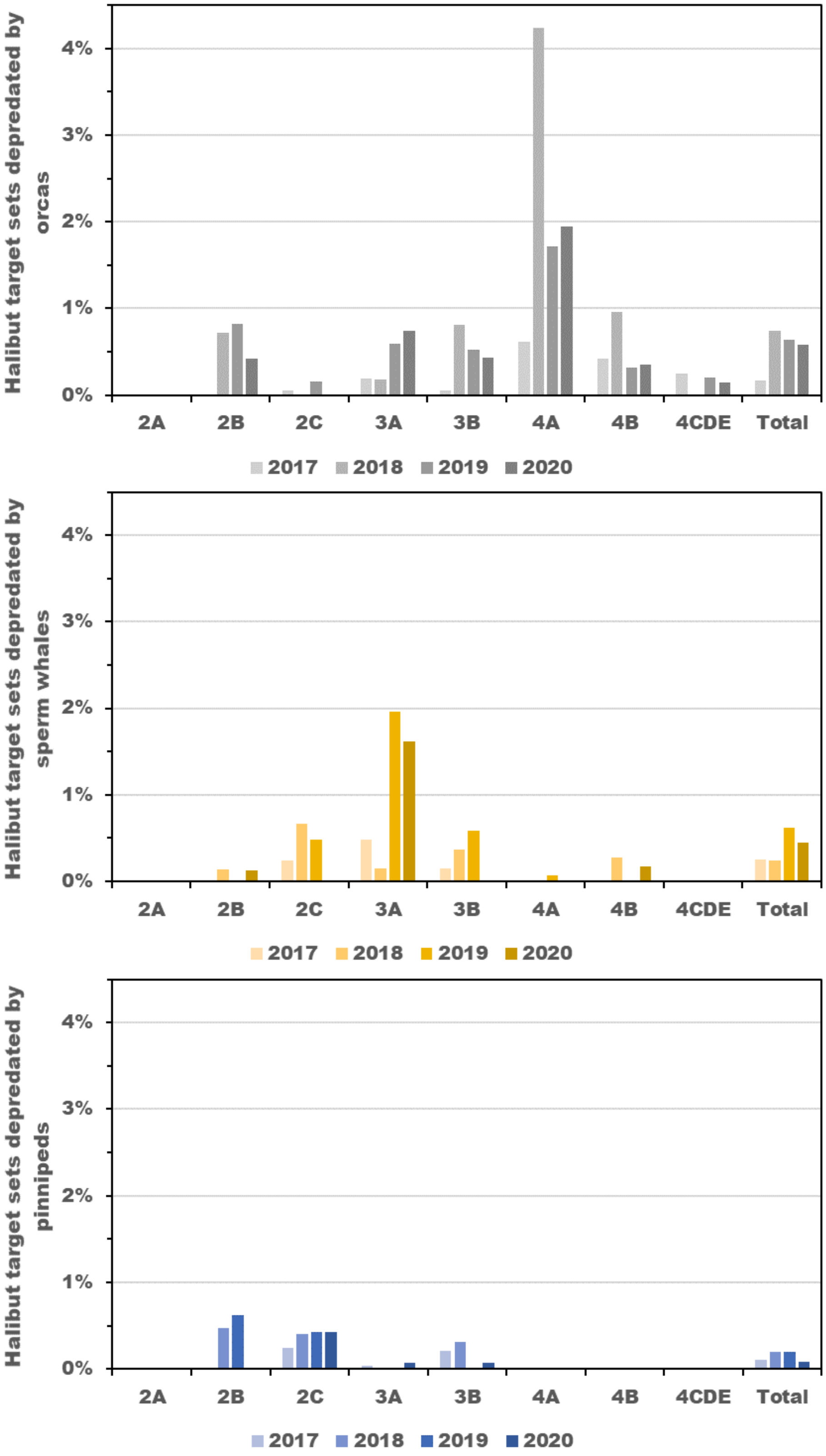


Figure 13. Percent of logbook-recorded sets with reported marine mammal depredation for orcas (upper panel), sperm whales (middle panel) and pinnipeds (lower panel) by IPHC Regulatory Area and year.

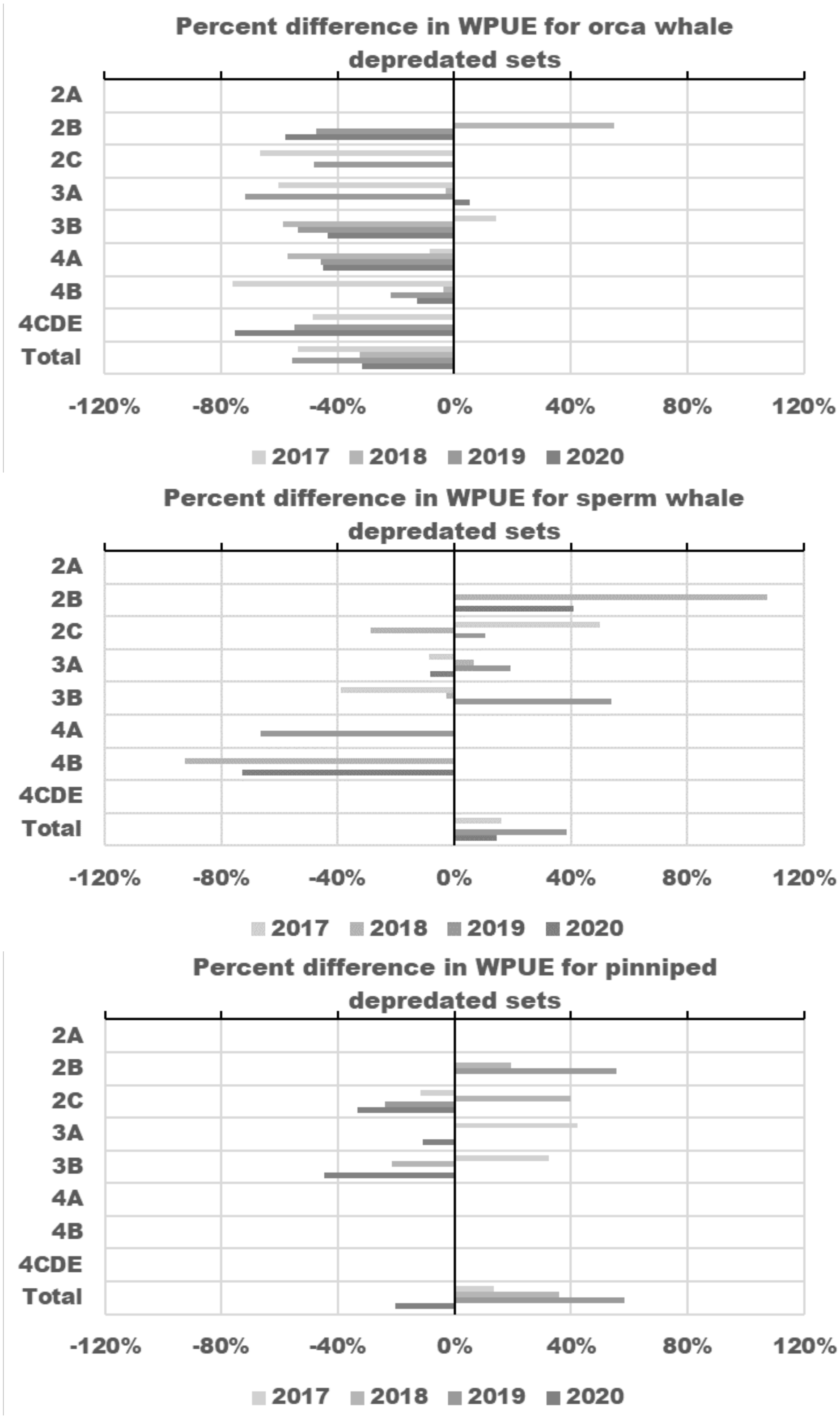


Figure 14. Percent difference in WPUE for logbook-recorded sets with reported marine mammal depredation for orcas (upper panel), sperm whales (middle panel) and pinnipeds (lower panel) compared to those with no reported depredation by IPHC Regulatory Area and year.

Model development

Multimodel approach

Creating robust, stable, and well-performing stock assessment models for the Pacific halibut stock has proven extremely challenging due to the highly dynamic nature of the biology, distribution, and fisheries (Stewart and Martell 2014). The stock assessment for Pacific halibut has evolved through many different modeling approaches over the last 30 years (Clark 2003; Clark and Hare 2006). These changes have reflected improvements in fisheries analysis methods, changes in model assumptions, and responses to recurrent retrospective biases and other evidence of model mis-specification and concurrent degradation of model performance (Stewart and Martell 2014). Perhaps the most influential of these changes was the transition from separate IPHC Regulatory Area-specific assessment models to a coastwide model in 2006, as the understanding of adult movement among areas was substantially updated by the results of the IPHC's extensive PIT-tagging experiment in 2003-2009 (Clark and Hare 2006; Webster et al. 2013). Some simulation studies have found that dividing a migratory population into several discrete assessment units tends to overestimate the total biomass (e.g., Li et al. 2014; McGilliard et al. 2014).

Although recent modelling efforts have created some new alternatives, no single model satisfactorily approximates all aspects of the available data and scientific understanding. Building on simpler approaches in 2012 and 2013, in 2014, the current ensemble of four stock assessment models, representing a two-way cross of short vs. long time series', and aggregated coastwide vs. AAF models was developed for the full assessment analysis and review in 2015 (Stewart and Martell 2016) and further improved in 2019 to accommodate sex-specific age composition data from the commercial fishery (Stewart and Hicks 2019b). AAF models are commonly applied when biological or sampling differences among geographical areas make coastwide summary of data sources problematic (Waterhouse et al. 2014). AAF models continue to treat the population dynamics as a single aggregate stock, but fit to each of the spatial datasets individually, allowing for differences in selectivity and catchability of the fishery and survey among regions. In addition, AAF models more easily accommodate temporal and spatial trends in where and how data have been collected, and fishery catches have occurred. This is achieved through explicitly accounting for missing information in some years, rather than making assumptions to expand incomplete observations to the aggregate coastwide level. Both aggregating the data into a single series and approximating spatial dynamics via AAF approaches may be useful under some circumstances; however, there is no clear best-performing configuration under all conditions. Not surprisingly, models that most closely match the biology, which is only known under simulated conditions, tend to perform the best (Punt et al. 2015).

To capture the structural uncertainty inherent among the Pacific halibut stock assessment models, it is necessary to use multi-model inference, here referred to as an 'ensemble' of models (e.g., Ianelli et al. 2016; Karp et al. 2018; Stewart and Martell 2015). The ensemble approach, applied in many fields in addition to fisheries (Du 2014; Hamill et al. 2012), recognizes that there is no "perfect" or "true" assessment model, and that a robust risk assessment can be best achieved via the inclusion of multiple models in the estimation of management quantities and the uncertainty about these quantities (Stewart and Martell 2015). This stock assessment is based on the approximate probability distributions derived from an ensemble of models, thereby incorporating the uncertainty within each model as well as the uncertainty among models. This approach reduces potential for abrupt changes in management quantities as improvements and additional data are added to individual models (Stewart and Hicks 2018), and provides a more realistic perception of uncertainty than any single model, and therefore a stronger basis for risk assessment.

The current ensemble explicitly captures two critically important dimensions of uncertainty: how the time-series data are used via short and long models, and how the spatial information is treated in the models via data aggregation to the level of Biological Regions treated as separate fleets (AAF) or to the coastwide level. Inclusion of these sources of structural uncertainty results in wider confidence intervals than are commonly seen in single-model stock assessments (Stewart and Hicks 2019a). More detail on how the models are weighted and integrated can be found in the Ensemble section below.

Structural rationale

Consistent with analyses since 2015, this stock assessment is implemented using the generalized software stock synthesis (Methot and Wetzel 2013a), a widely used modeling platform developed at the National Marine Fisheries Service. This platform allows for a wide range of structural choices with regard to biology and growth, catchability, selectivity, spatial processes, stock-recruitment dynamics as well as error distributions and integrated projections. A benefit of using this code is that it is well documented, and the inputs and output formats are standardized (Methot Jr et al. 2021a), regardless of model configuration, allowing easy interpretation of model files and rapid evaluation of the results without re-running the fitting algorithm using the r4ss package (Taylor et al. 2021) implemented in the R programming language (https://cran.r-project.org/).

A primary structural stock assessment model choice is whether or not to model growth explicitly (and often parametrically) or empirically. Many stock assessments assert/estimate a growth function of some type and rely on this growth function to translate between numbers and biomass for model calculations. This approach has the benefits of allowing direct fitting to observed length observations, interpolating and/or extrapolating predictions for years where direct observations may be missing, as well as direct inclusion of the potential effects of selectivity at length on the observed data. The cost of such an approach is that growth can be an extremely complex process, varying over time, space and by cohort (via density dependence). When there is appreciable growth variability, a great deal of complexity may be required to adequately model this population process, even before sampling and selectivity issues have been addressed. Failure to account for this type of variability can lead to poor fits to composition data, potentially biasing the assessment results (Maunder et al. 2015, and subsequent special issue papers).

The Pacific halibut stock assessment models, like many other stock assessments with relatively complete age and size information, take a simpler approach to growth by using empirically derived weights-at-age. The empirical weight-at-age approach has the benefit of reducing complexity with regard to growth modelling but has several costs in other modelling areas. These include the need for more complexity in modelling selectivity, particularly where some of the selectivity process may be a function of size rather than age alone. This is the case for Pacific halibut, where the interaction of changes in size-at-age, gear selectivity that is likely at least partially a function of fish size, and a minimum size limit thus requires the treatment of selectivity-at-age as a time-varying process (Stewart and Martell 2014). However, the treatment of selectivity as time-varying appears to be a necessity for Pacific halibut even if treated as a function of size; static selectivity for a spatially aggregated model in the face of changes in availability was identified as a primary contributor to severe historical retrospective patterns (Stewart and Martell 2014).

There are relatively few examples of stock assessments used for management purposes that are explicitly spatial: modelling movement among areas, distributing recruitment events, and tracking spatial variability in biological characteristics (e.g., McGilliard and Palsson 2021; Stewart et al. 2009). Most such cases rely on low rates of movement to allow for estimation of recruitment distribution among areas. More frequently assessments either aggregate the available data across spatial heterogeneity (preferably weighting appropriately such that the aggregate information reflects the underlying distribution), or retain separate data series representing spatial areas, but fit to them in the context of a single instantaneously mixing population model (the AAF approach). These methods for dealing implicitly with spatial dynamics are by necessity gross approximations, with performance properties specific to a particular application that are unknown, and almost certainly depend on the true underlying processes. Some simulation studies have shown that fisheries operating in different areas with differing selectivity schedules can be reasonably approximated by an AAF approach (e.g., Waterhouse 2014). Other studies have found acceptable performance of AAFs when simulating actual spatial variability (e.g., Hurtado et al. 2014, McGilliard et al. 2014); however additional studies have found that combining spatial data into weighted aggregates also performs acceptably and may be more stable than more complex AAF approaches (Punt et al. 2015, Li et al. 2015). A primary conclusion from simulation-based studies is that if the true underlying process is well-represented, then models reflecting these dynamics tend to perform well (Goethel and Berger 2017). Unfortunately, in the case of Pacific halibut it is not clear whether aggregated or AAF models might be the best choice as neither approach accurately represents the complex spatial dynamics.

The choice of how long a time-series to model generally represents a compromise among: data availability, data quality, model complexity, and technical convenience (e.g., data preparation and model convergence times). As assessment model time series' are extended to include more historical data, commonly the quality of those data becomes increasingly lower as standardization of sampling programs has a greater likelihood of having changed appreciably. In the case of Pacific halibut, fishery-independent survey information has been reasonably comprehensive since approximately 1997, and sufficient to support the recently developed geostatistical model since 1993 (Webster 2018). Current fishery sampling approaches have also not changed dramatically over the same period. The completeness of this time period with regard to data availability was one of the primary incentives for stock assessment models used by the IPHC since 2006 to begin the modelled period in 1996. Notable differences prior to that period included the transition in the survey and fishery from “J” to circle hooks, variable and much less comprehensive survey coverage, lack of access to raw historical fishery data (ages, catch rates, etc.), and many others. The costs of using only a relatively short time-series include a lack of integration between harvest strategy calculations derived from full historical period, a lack of perspective on recent trends, the need for careful treatment of initial model conditions, inability to estimate some parameters, and increased sensitivity to additional data, as each year represents a greater fraction of the total information available in the model. These trade-offs prompted the development of the first long time-series model in 2013, with the recognition that neither the short or long time-series approach was clearly superior, and that differences in the results reflected a meaningful source of uncertainty in the assessment results.

All of the halibut models considered here treat male and female halibut separately. Like many broadcast spawning fishes, there is a basic assumption that spawning is likely to be limited primarily by female spawning output and not by male abundance (at least over a reasonable range of sex-ratios; this is generally not a concern except for cases such as some crab stocks where fishery mortality may operate primarily on males). If the sex-ratio could be expected to be stable over time, it might be reasonable to structure assessment models without regard to sex and/or just assume half of the mature biomass represented females. However, for Pacific halibut, highly dimorphic growth interacting with gear selectivity for larger fish, and a fishery in which there are strong incentives to target the larger females (due to the minimum size limit and graduated price structure) results in sex-ratios of the catch and of the landings skewed largely toward females. Historical modelling suggested that the potential for a static assumption regarding sex-ratio could lead to a highly biased interpretation of stock status and that females and males are best modelled separately.

In aggregate, these considerations led to the choice of four stock assessment models during the 2014 assessment process: a two-way cross of: coastwide vs. AAF data structuring, and long vs. short time-series. Each of these models explicitly treated male and female halibut separately and employed empirical weight-at-age rather than an explicit growth function. All models fit to both fishery and survey index trends and age compositions and allowed for temporal variability in selectivity and catchability. Additional alternative modelling approaches were considered, including a simple surplus production model and a Virtual Population Analysis model. Both of these approaches suggested that recent removals and stock trends were on a similar scale to the four models included in that assessment (Stewart and Martell 2015) but presented sufficiently substantial issues in interpretation or application to the management process that they were not formally included in the final stock assessment.

General model configuration

There are a number of basic technical settings and features that are common to all four stock assessment models described here. This section provides an overview, which is supplemented by a description of specific individual model details below.

The stock synthesis software separates inputs into several files read in prior to model estimation including the primary data file, the primary control file (including parameter setup and estimation switches), the weight-at-age file, the forecast file (including settings for reference point calculations), and the starter file (including some general estimation and reporting switches and settings). Each of these input files for each of the four stock assessment models described here are included in the background documents, along with the primary report file of estimated and derived quantities and the directory of summary and diagnostic figures created by r4ss (see Appendix A). Note that not all automatically created diagnostic material is relevant to the model configurations employed here.

These models were configured to make use of relatively standard population structuring. There were no seasonal dynamics, and catches were assumed to be removed halfway through the year via Pope's approximation. This approach does not require iterative estimation of fleet- and year-specific fishing mortality rate parameters (often reducing model run times) and should reasonably approximate the dynamics unless fishing mortality rates are extremely high. Catches were input in thousands of pounds (net weight; head-off and gutted, approximately 75% of round weight), so that the mean weight-at-age inputs were in net pounds and the numbers-at-age are tracked in thousands of individuals. Population dynamics contain ages 0-30, and female and male halibut are modelled separately in the underlying dynamics.

The input data were partitioned via a fleet structure of: the directed fishery (by area in the AAF models), discard mortality from the directed fishery, non-directed discard mortality ('bycatch'), recreational, subsistence, and survey (FISS; by area in the AAF models). Table 14 summarizes the data and key features of each model (note that all changes from the 2021 model are described in greater detail below). Age data were partitioned by sex (the vectors for each year contain females, then males, such that the sex-ratio is inherently included in the age compositions), where this information was available and assigned the appropriate ageing method in the data file (see section above). Where few fish contribute to the 'tails' of the age distributions for each fleet and year combination, the model was set to automatically aggregate observations and predictions at each of the low and high ages with proportions less than 0.1%. This choice avoids large vectors of zeroes in the multinomial calculations. The model was also set up to add a very small constant (0.0001) to all age proportions in order to stabilize the computation.

Table 14. Comparison of structural assumptions among models.

	Model			
	Coastwide Short	Coastwide Long	AAF Short	AAF Long
Modelled period[1]	1992+	1888+	1992+	1888+
Data partitions	N/A	N/A	Regions 2, 3, 4, 4B	Regions 2, 3, 4, 4B
Directed Fishery fleets	1	1	4	4
Other fishing fleets	4	4	4	4
Survey fleets	1	1	4	4
Fishery CPUE (weight)	1992+	1907+	1992+	1907+, 1915+, 1981+, 1981+
Fishery age data years	1992+	1935+	1992+	1935+, 1935+, 1945+, 1991+
Survey CPUE (numbers)	1993+	1977+	1993+, 1993+, 1997+, 1997+	1977+, 1977+, 1997+, 1997+
Survey age data years	1993+	1963+	1993+, 1993+, 1997+, 1997+	1965+, 1963+, 1997+, 1997+
Weight-at-age	Aggregate	Aggregate	Areas 2, 3, 4	Areas 2, 3, 4
Female *M*	Fixed at 0.15	Estimated	Estimated	Estimated
Male *M*	Estimated	Estimated	Estimated	Estimated
Stock-recruit relationship	B-H	B-H	B-H	B-H
Initial conditions estimated	R_{init} *N*-at-age: 1-19	R_0, *N*-at-age: 1-29	R_{init}, *N*-at-age: 1-19	R_0, *N*-at-age: 1-29
Environmental regime effects on recruitment	No	Estimated	No	Estimated
Steepness (*h*)	0.75	0.75	0.75	0.75
$\sigma_{\text{recruitment deviations}}$	1.0	0.54	0.80	0.5
Survey selectivity	Asymptotic, by sex	Asymptotic, by sex	Domed, by sex (R2, R3) Asymptotic, by sex (R4, R4B)	Domed (R2, R3), Asymptotic (R4, R4B)
Fishery selectivity	Asymptotic, by sex	Asymptotic, by sex	Domed, by sex (R2, R3) Asymptotic, by sex (R4, R4B)	Domed, by sex (R2, R3) Asymptotic, by sex (R4, R4B)
Scale of male fishery selectivity	Estimated, time-varying	Estimated, time-varying	Estimated, time-varying	Estimated, time-varying
Non-directed discard selectivity	Domed	Asymptotic	Domed	Domed
Recreational selectivity	Asymptotic	Domed	Domed	Domed
Discard selectivity	Domed, by sex	Domed, by sex	Domed, by sex	Domed, by sex
Subsistence selectivity	Mirrored to recreational	Mirrored to recreational	Mirrored to recreational	Mirrored to recreational

[1] Mortality estimates for 2022 were projected based on adopted IPHC limits.

All growth specifications in the control file were bypassed in order to use the empirical weight-at-age approach; therefore, the settings in the control file and the results included in model outputs related to these settings are not meaningful (this includes length-at-age, weight-at-length, and maturity-at-length; these are all integrated directly in the weight-at-age inputs). The weight-at-age file also included a matrix of spawning output-at-age representing the product of annual weight-at-age (a matrix) and the static vector of maturity-at-age (Stewart and Webster 2022).

For all estimated parameters (except temporal deviations), uniform priors were implemented, with bounds sufficiently wide to avoid maximum likelihood estimates falling on or very near a bound, unless the bound was structurally logical. Table 15 summarizes the counts of estimated parameters in each model. Natural mortality was allowed to differ by sex, with the value for male halibut estimated in all four models, and the value for females in all but the short coastwide model. Treatment of both the stock-recruitment relationship and the initial conditions at the start of the modelled time-series differed among the four models and are described below.

The double-normal selectivity parameterization is used in all four models, as it represents a flexible, but still parametric approach that can easily be made time-varying via just one or two parameters with annual deviations. There are more flexible nonparametric selectivity options, but these generally require all the parameters to vary over time, creating a substantial increase in complexity. The double-normal selectivity can be easily configured to be either asymptotic or dome-shaped, by adjusting the width of the peak and/or descending slope and final selectivity parameters. It also includes an option for male selectivity to be offset from female selectivity, based directly on the parameters of the selectivity curve (females from males), such that time-varying selectivity for one sex can be mapped into variability for both sexes without estimating a second set of parameters. The double-normal was implemented for all model fleets, with at least the ascending limb of selectivity (ascending width and peak parameters) allowed to vary over time for all four models (described further below).

As has been the case in all recent halibut models, the catch-per-unit-effort index derived from the directed halibut fishery is included in each of the models, but the catchability is allowed to vary over time, except in a few cases where there was no improvement in model fit by allowing temporal variability or where iterative tuning of the degree of interannual change suggested no meaningful variation. In principle, there are many factors which can create changes in the proportionality of the catch-rate in a fishery with the underlying population. The most obvious of these are abrupt changes in fishing methods, such as the change from “J” to circle-hooks in 1984. This type of change was accommodated (in the long time-series models) via an unconstrained deviation on catchability in that year (effectively a separate *q* for the two parts of the time series). Beyond abrupt changes, there are many factors that can ‘drift’ over time, but may not be so obvious, including technological improvements, changes in spatial areas or times of year being fished, etc. This type of change suggests a random walk in catchability, which was the approach taken in all four models here. To implement this, a catchability parameter was estimated for the first year for which index data were available, and then a deviation (from the previous year’s value, not the mean) was estimated for each subsequent year of the time-series.

The annual catchability deviations were constrained by a single σ for each fleet. The iterative tuning algorithm for identifying the internally consistent values for each σ is described below along with other changes for 2021.

Table 15. Comparison of estimated parameter counts among models.

	Model			
	Coastwide Short	Coastwide Long	AAF Short	AAF Long
Static				
Female *M*	--	1	1	1
Male *M*	1	1	1	1
Log(R_0)	1	1	1	1
Initial R_0 offset	1	--	1	--
Environmental link coefficient	--	1	--	1
Fishery catchability	1	1	4	4
Survey catchability	1	4	--[1]	4
Fishery selectivity	5	5	21	20
Discard selectivity	8	7	5	5
Non-directed discard selectivity	4	2	3	3
Recreational selectivity	5	6	5	6
Survey selectivity	5	5	21	18
Total static	32	34	63	64
Time-varying				
Recruitment deviations[2]	54	168	54	168
Fishery catchability deviations	--	111	116	218
Fishery selectivity deviations	78	175	244	568
Survey selectivity deviations	84	90	206	260
Total deviations	216	544	620	1,214
Total	248	578	683	1,278

[1] The analytic solution is used for these catchability parameters.

[2] Includes initial age structure and five uninformed forecast years (the latter only included here such that counts will match that reported in model output).

In all models, fit to the age data used a multinomial likelihood with initial input sample sizes based on the bootstrap results described above, subsequently adjusted down via a multiplicative scalar for each fleet in the control file (more discussion below). Indices of abundance from both the FISS and commercial fishery (by area in the AAF models) were fit using a log-normal likelihood and input log(*SE*)s based on the space-time modelling. Survey indices were fit in numbers of fish to avoid converting numbers to weights in the data and then weights back to numbers in the model predictions (as informally recommended by the Scientific Review Board in 2014). Weight-per-unit-effort is the native scale for the fishery indices based on logbook records.

Using the method first developed for the 2015 assessment, discard mortality, bycatch and recreational selectivity are estimated, but the age composition data are down-weighted to avoid imparting any significant information on recruitment strengths from these uncertain and potentially non-representative data sets. In this way, the data that are available serve as an informative ‘prior’ on the selectivity for each of these fleets, and therefore propagate some uncertainty associated with selectivity estimation, but do not strongly inform other model parameters and population dynamics estimates.

Discards in the directed commercial fishery are treated as a separate fleet in each model. This approach was taken for several reasons: discard rates may be a function of spatial fishing effort and not simply contact selectivity as is often assumed in stock assessments - there has been little relationship between the magnitude of discards and the magnitude of commercial landings when this has been evaluated for previous reviews. Further, modelling discards with a retention curve in the empirical weight-at-age approach within SS does not allow for separate mean weight-at-age vectors to be applied to landings and discards (which may differ significantly for younger ages due to the size limit). Sex-specific selectivity curves were estimated in each model informed by the observations from the sublegal fish captured by the setline survey. The selectivity was configured to be a double normal, with female halibut offset from male halibut to account for the dimorphic growth (the opposite of all other fleets), and the relative scale of females to males estimated directly. Both sexes were allowed to be dome-shaped, with differing descending limbs. Because the sublegal survey age data were already included in the likelihood as part of the survey age compositions, it would be a misrepresentation of the uncertainty to naively fit them again equally as part of the discard data set. Instead, previous analyses showed that down-weighting these data such that they had a very small input sample size had no appreciable effect on the model results but still allowed for the direct estimation of selectivity. This approach lends itself to direct inclusion of observer data on discarded halibut when sampling/expansion methods that are representative of the entire fleet become available.

Bycatch and recreational selectivity curves were also allowed to be dome-shaped given the relative frequency of younger halibut in the observed distributions. Where descending limb parameters were estimated to be at the upper bounds, these parameters were fixed (making the curves asymptotic) to avoid any negative behavior during minimization and approximation of the variance in model quantities via the Hessian matrix. Since the 2019 assessment, sex-specific age composition data for the recreational fishery has become available (Stewart and Webster 2022), and so additional offset parameters were added to allow for sex-specific selectivity as in the treatment of the discards. Because of the down-weighting of the data for these series, and the unknown or potentially poorly spatially representative nature of the data themselves, no attempt was made to allow these selectivity curves to vary over time.

The presence of both observation error (in the indices and age composition data) and process error (in fishery catchability and selectivity for the survey and fishery) creates a challenge for standard weighting and tuning practices employed in many assessment models. Specifically, if process error is not modelled (and/or a fixed value is asserted), the input sample sizes (and sometimes index variances) can be relatively easily iteratively tuned or estimated (Maunder

2011). This approach is useful for reducing the potential effects of outliers, lack-of-fit, or model misspecification with regard to composition data (Francis 2011). At the other extreme, if the observation error is assumed to be known (and assigned a fixed value), then the degree process error can be estimated via random effects, or iteratively tuned using a maximum likelihood-based approximation (the ‘Thompson and Lauth method’; Annex 2.1.1 in Thompson and Lauth 2012). When data are sufficient, both components can be iteratively or by more statistically rigorous means estimated simultaneously (Thorson 2019; Thorson et al. 2016).

The general goal for the treatment of process error in selectivity and catchability and observation error in the data is to first reduce clear signs of bias to the degree possible and then to achieve internal consistency among error distributions and sample sizes/variances. In all four models developed here, the initial input sample sizes, for 2022 derived from the bootstrapping analysis described above were considerably larger than commonly applied weighting for stock assessment models would suggest (Table 11-13). These values were iteratively reduced based on evaluation of three considerations: the relative magnitude of the standardized Pearson residuals, comparison of the input value for each fleet with the harmonic mean effective sample size which is an unbiased estimator for a set of independent multinomial samples (Stewart and Hamel 2014), and the scaling suggested by the Francis (2011) method (as implemented in the r4ss package). For almost all fleets and all models, this approach led to a substantial reduction from initial sample sizes. In no cases were the input values increased from the maximum values derived via bootstrapping.

Starting from a small value for the input σ for each fleet and parameter combination where temporal variability was allowed, process error was increased until the tuned value was consistent with the degree of variability observed among the deviations (SE_{devs}^2) and the average uncertainty of the deviations themselves $\bar{\sigma}_{dev}{}^2$. This approach is very close to that outlined by Thompson and Lauth (2012) and is consistent with the preferred method for tuning this and other types of process error (such as recruitment deviations) in stock synthesis (Methot and Taylor 2011; Methot et al. 2019):

$$\sigma_{tuned} \sim \sqrt{SE_{devs}^2 + \bar{\sigma}_{dev}{}^2}$$

In addition to providing internal consistency, this approach makes intuitive sense: under perfect information the average variance of the deviations will be zero and the variability among the deviations will exactly match the process error, conversely under no information the variance of the deviations will be the input constraint. After initial process error tuning, the input sample sizes were adjusted downward until the weights suggested by the fit to the mean age over the time series were approximately equivalent to the input values (the “Francis method’; Francis 2011). There were only minor changes to the tuned σ values required after iteration of the input sample sizes, suggesting the two processes were relatively separable and stable; further there were only minor changes in the process error variances in this assessment relative to the 2019 assessment despite the revised input sample sizes.

As a final model-building step, models were regularized via adjusting parameterizations through removing and/or fixing selectivity parameters that consistently remain stuck to bounds or are not contributing to the likelihood in a meaningful way (<1% correlation with other model parameters). This regularization does not include forecast recruitment deviations, which are expected to be uncorrelated with other model parameters (and the objective function), but are 'estimated' in order to appropriately propagate the uncertainty in recent recruitments into forecasts.

The tuning approach for the stock-recruitment relationship was very similar, ensuring that the input σ governing recruitment variability was consistent with the observed variability and variance estimates; the calculation for this tuning is automated in the r4ss package, and the output was used as a guide for the scale of the bias correction, including ramps to and from the peak value consistent with the information content of the data and variability in the deviations observed in the output. This step is important for recruitment variability as it also provides for a better approximation for the bias correction in recruitment deviations (Methot and Taylor 2011) in the 'main' or best informed period of the time-series of recruitments. Again here, after initial tuning, little change was observed across alternative models (or from the 2019 results).

In the end, this tuning process provides a model that is internally consistent: the error distributions are commensurate with the fit to the data and the degree of process error is consistent with the signal (information content) in the data. Importantly, accounting for process error in selectivity was the primary solution for historically observed retrospective patterns in the Pacific halibut stock assessment models (Stewart and Martell 2014). Tuning diagnostics and results specific to each model are provided below.

Coastwide short

The initial conditions for a model starting after an extensive historical fishery and appreciable recruitment variability must be structured to avoid simple assumptions that may have strong effects on the subsequent time-series. For the coastwide short model the initial conditions included estimating the population numbers at age 1-19 in the first year of the model (1992 after extension of the time-series; see below). Since the age data available for the initial year were aggregated at age-20 (due to the historical use of the surface ageing method), there was no specific information on additional individual year-classes. To accommodate a non-equilibrium value in the plus group, an offset to initial equilibrium recruitment (via a single time 'block') was also estimated. The effect of these two approaches was to essentially decouple the numbers-at-age at the beginning of the time-series from any equilibrium assumptions.

As in previous assessments, the coastwide short model employed a Beverton-Holt stock recruitment relationship with estimated equilibrium recruitment level (R_0) setting the scale of the stock-recruit relationship. Steepness (h) was fixed at a value of 0.75 for this and all other models, an assumption that has been explored extensively in previous assessments. Fixing steepness, but iteratively solving for the internally consistent level of recruitment variability generally does not have a large effect on year-class strengths where data are informative, but does have very strong effects on direct estimates of Maximum Sustainable Yield (Mangel et al. 2013); however, this quantity is not of specific interest for the Pacific halibut assessment. A summary of the number of estimated parameters contributing to each aspect of the model is provided in Table 15.

Age-based selectivity for female halibut for both the FISS and commercial fishery was estimated using the double normal, forced to be asymptotic once it reached peak selectivity. This required two parameters: the ascending width of the curve and the age at which the peak selectivity is reached. Both parameters are allowed to vary over time with a random walk of annual deviations. These deviations were initiated in the first year for which age composition data were available, and no deviation was estimated for the terminal year (2022), because there were no data yet in the model to inform that deviation (it will be estimable when the 2022 data are added for the final assessment). This means that the actual mortality in 2022, when available, may have a different effect than initial projections. Male selectivity for the survey was estimated via offsets to the female ascending width and peak parameters, and a third parameter defining the scale of male selectivity relative to that for females. In the coastwide short model, with fixed female natural mortality and direct overlap between all years of fishery and survey age data, the male offset parameters for the fishery have been estimated in recent assessments. These parameters have been informed by the weak information on sex-ratio included the sex-aggregated age data. In aggregate, there were five estimated base parameters each for the survey and fishery and annual deviations on the ascending limb parameters (Table 15).

As in the 2015 and 2019 assessments, the scale of male selectivity for both the survey and fishery were allowed to vary over time as a random walk. With only sex-aggregated commercial fishery age compositions prior to 2017, it is not clear how strongly the temporal variability in the scale of male selectivity is informed (and potentially how correlated it would be with female natural mortality, which is fixed in this model). However, the addition of time-varying deviations on the scale parameters was found to improve the residual patterns in previous assessments for the fit to the fishery age-data and has not shown signs of unreliable estimation over sensitivity and alternative model runs.

Coastwide long

Initial conditions for the coastwide long time-series model include the initial age structure and a long period of uninformed recruitments with the model period beginning in 1888 and the first age data available for 1935 (Table 14); therefore, there was a substantial ‘burn in’ for recruitment variability prior to any data. The treatment of the stock-recruitment function in the coastwide long model was substantially different from that of the coastwide short model. Consistent with historical IPHC analyses (Clark and Hare 2002a, 2006) and previous stock assessments, the coastwide long model allowed for the possibility that recruitment variability is correlated with the regimes of the Pacific Decadal Oscillation (PDO; Mantua et al. 1997). To implement this approach, a Beverton-Holt relationship was used, parameterized with an estimated value for the equilibrium recruitment level (R_0) parameter, and a fixed value of steepness (h) of 0.75. The annual average of the PDO index was converted to a binary indicator (PDO_{regime}) where productive regimes (e.g., 1977-2006) were assigned a value of 1.0, and poor regimes (e.g., 1948-1976) a value of 0.0 (Stewart and Webster 2022). These regimes were linked to the scale of the stock-recruit function via an adjusted equilibrium level of recruits (*R₀'*) based on an estimated coefficient (*β*) creating an offset to the unadjusted value:

$$R_0{}' = R_0 * e^{\beta * PDO_{regime}}$$

The adjusted equilibrium recruitment value was then used in the stock-recruit function with bias-corrected annual deviations:

$$R_y = f\left(SB_y, R_0{}', SB_0, h\right) * e^{r_y - \frac{\sigma^2}{2}}$$

This approach changed since 2015 (but see alternative sensitivity analyses below). This parameterization has the desirable property that if there is no correlation between the putative environmental index and underlying mean recruitment, the *β* parameter to be estimated at a value of 0.0 and the recruitment estimates will be unaffected. In that case R_0' is simply equal to R_0. As was the case for the coastwide short time-series model, fixing steepness precludes the naïve use of *MSY* estimates.

The approach to selectivity in the coastwide long model was identical to that in the coastwide short model, except that the annual scale of male selectivity was only estimable after adding the sex-ratio information beginning in 2017 (see changes from 2021 below) Selectivity deviations on the ascending limb parameters of the fishery and survey series were initiated in the first year for which age composition data were available for both the fishery (1935) and the survey (1963).

AAF short

The AAF short model was configured very similarly to the coastwide short model. The most notable difference was in the treatment of selectivity for the survey and fishery in Biological Regions 2 and 3: these were allowed to be dome-shaped relative to the coastwide population dynamics. Implementing dome-shaped selectivity for these four model fleets requires the addition of a third selectivity parameter defining the width of the descending limb. This additional parameter was not allowed to vary over time.

Another difference between the short time-series models was in the treatment of the scale of male selectivity for the fishing fleets in each of the four areas. Similar to the coastwide long model, the three parameters defining the annual male offset to female selectivity for the commercial fishery in each area were only estimable beginning with the 2017 sex-ratio data. Temporal variability in selectivity parameters occurred over a slightly longer range of years in the AAF short model, as there were Region-specific survey data available for the entire time-series from Biological Regions 2 and 3.

Finally, unlike in the 2021 and earlier assessments (as described below), the preliminary 2022 short AAF model estimates female *M*. Likelihood profiles (see below) suggested a defined and reasonable minima somewhat closer to the long coastwide and AAF models than the previously assumed value of 0.15.

AAF long

The only structural differences between the AAF long and AAF short models were the years over which deviations in recruitment, selectivity and catchability are estimated. The AAF long model treated the stock-recruitment function in the same manner as the coastwide long model, including the PDO as an estimated covariate to equilibrium recruitment.

Changes from 2021

In the intervening period between the last full stock assessment analysis and review in 2019 and this preliminary analysis for 2022, the length and information content of the data sets has grown, and new information, such as the bootstrapping results (described above) has become available. Changes to specific data sets have been documented in the recent assessments and their effects evaluated singly in each year (Stewart and Hicks 2021; Stewart and Hicks 2022). Key changes for 2021 included:

1) Extending the time series to include projected mortality based on limits adopted for 2022 (IPHC 2022),
2) updating to the newest stock synthesis software version (3.30.19; Methot Jr et al. 2021a),
3) expanding the treatment of natural mortality (*M*) to include an informative prior and increased values at the youngest ages based on meta-analyses,
4) improving the basis for data weighting via use of bootstrapped effective sample sizes based on the FISS and fishery sampling programs as model inputs (rather than the raw number of sets/trips),
5) re-tuning the process and observation error components of these models to achieve internal consistency within each,
6) allowing for interannual variability in the sex-ratio of the commercial fishery selectivity,
7) and exploring whether female *M* in the short models was estimable (male *M* and *M* for both sexes in the long models was already estimated).

The sequential effects on the model results of each of these changes are described below as a ‘bridging’ analysis from the 2021 stock assessment.

Extending the time-series

In order to provide for transparent comparisons from this preliminary stock assessment through the final results for 2022, the initial step in this analysis was to extend the modelled time-series to 2022, using the projected mortality associated with the limits set by the IPHC (IPHC 2022). Weight-at-age was assumed to remain constant from 2021 to 2022; however, it will be updated when new data become available. No other information was needed for this single year projection and all model results and parameter estimates remained unchanged relative to the final 2021 stock assessment.

Software version update

The Pacific halibut stock assessment has updated to newer versions of the stock synthesis software (Methot and Wetzel 2013a; Methot and Wetzel 2013b) as new features have been added, and in order to avoid major changes as input/output changes have evolved over time. The 2019 stock assessment was implemented in version 3.30.13 (Methot et al. 2019), which was updated to 3.30.15 (Methot Jr et al. 2020b) for the 2020 stock assessment in order to utilize the automatic calculation of variance and covariance for dynamic unfished stock size, a feature that was unavailable in previous years. For 2021, version 3.30.17 (Methot Jr et al. 2021b) was used, but the results were unaffected as there were no changes made that were related to any of the features used for Pacific halibut. Similarly, for 2022 the models were updated to version 3.30.19, but the results were identical to those produced under the previous version. For simplicity, this step has been omitted from the bridging figures below.

Treatment of M

As described above an informative prior was developed for use on both male and female Pacific halibut *M*. In addition, elevated *M* for ages 0-2 was also introduced to the assessment models in order to facilitate an in-depth exploration of the PDO as a covariate with recruitment strengths. The change did not affect the two short time-series models that had *M* fixed at 0.15 and had only a tiny effect on those models that estimated female *M*, slightly reducing the MLE for the long coastwide and long AAF models (Figure 15-17). This affect was consistent with the mode of the density for the informative prior slightly lower than the point estimates from the models, but with a large variance (Figure 19).

Addition of elevated *M* at ages 0-2 also had little effect on model estimates of spawning biomass (upper panels, Figure 15-17). The exception was the AAF long model (Figure 18), which was quite sensitive to any change affecting the historical time series (see discussion of convergence and likelihood profile sections below). In contrast to the spawning biomass time-series, the absolute estimates of recruitment increased substantially for all four models in order to generate numbers of fish at ages 3+ consistent with previous model fits (lower panels, Figure 15-17).

Data weighting

The next step in the bridging analysis was to replace the previously used input sample sizes (the number of samples contributing to the FISS and fishery age composition data) with the bootstrapped maximum effective sample sizes described above. The effective sample sizes were also tuned (as described above) during this step based on the calculated Francis weights and the magnitude of observed residuals. There were no clear directional patterns in the results of this change and changes to the estimated time-series were minor (Figure 15-17).

After revising and tuning the bootstrapped input sample sizes, process error variances were again iteratively tuned along with another iteration of the data weighting to ensure that all model configurations were internally consistent. Despite discovery and correction of an error in the implementation of time-varying catchability (leaving out several years from the block design) during this step changes were again relatively minor when compared to the uncertainty estimates and other bridging steps (Figure 15-17).

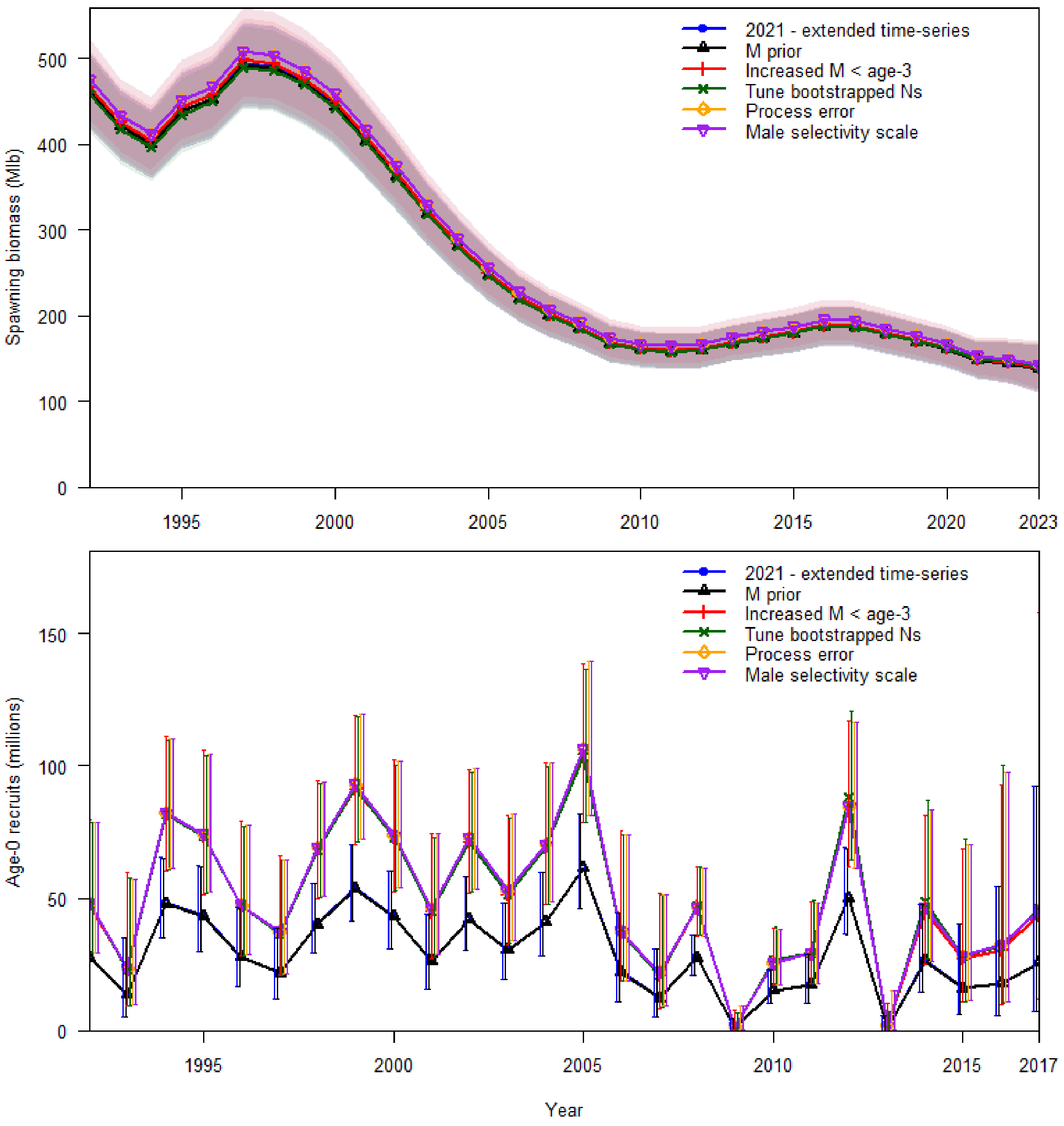


Figure 15. Comparison of estimated biomass (upper panel) and recruitment time series (lower panel) over sequential changes from the 2021 to preliminary 2022 coastwide short models.

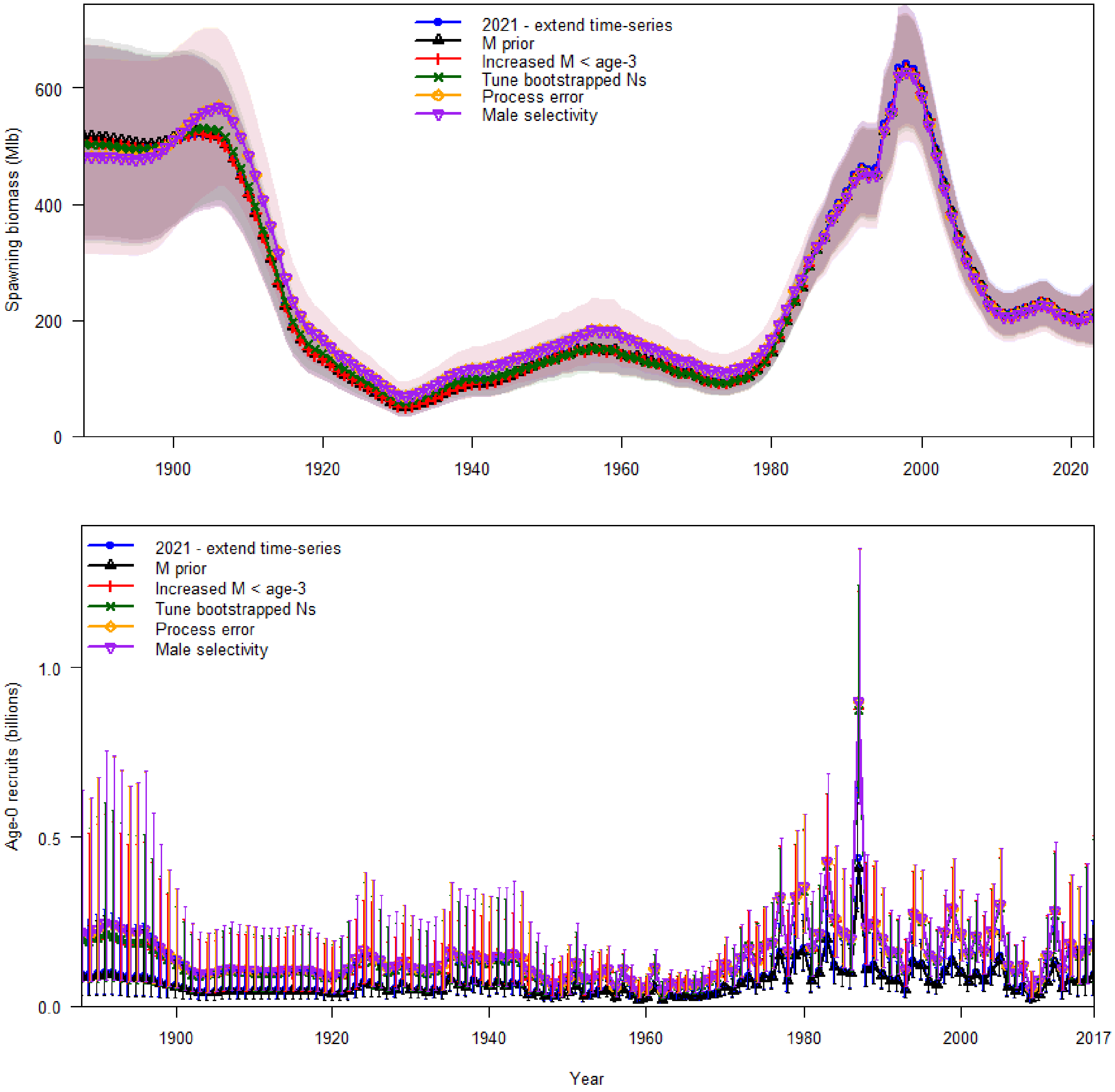


Figure 16. Comparison of estimated biomass (upper panel) and recruitment time series (lower panel) over sequential changes from the 2021 to preliminary 2022 coastwide long models.

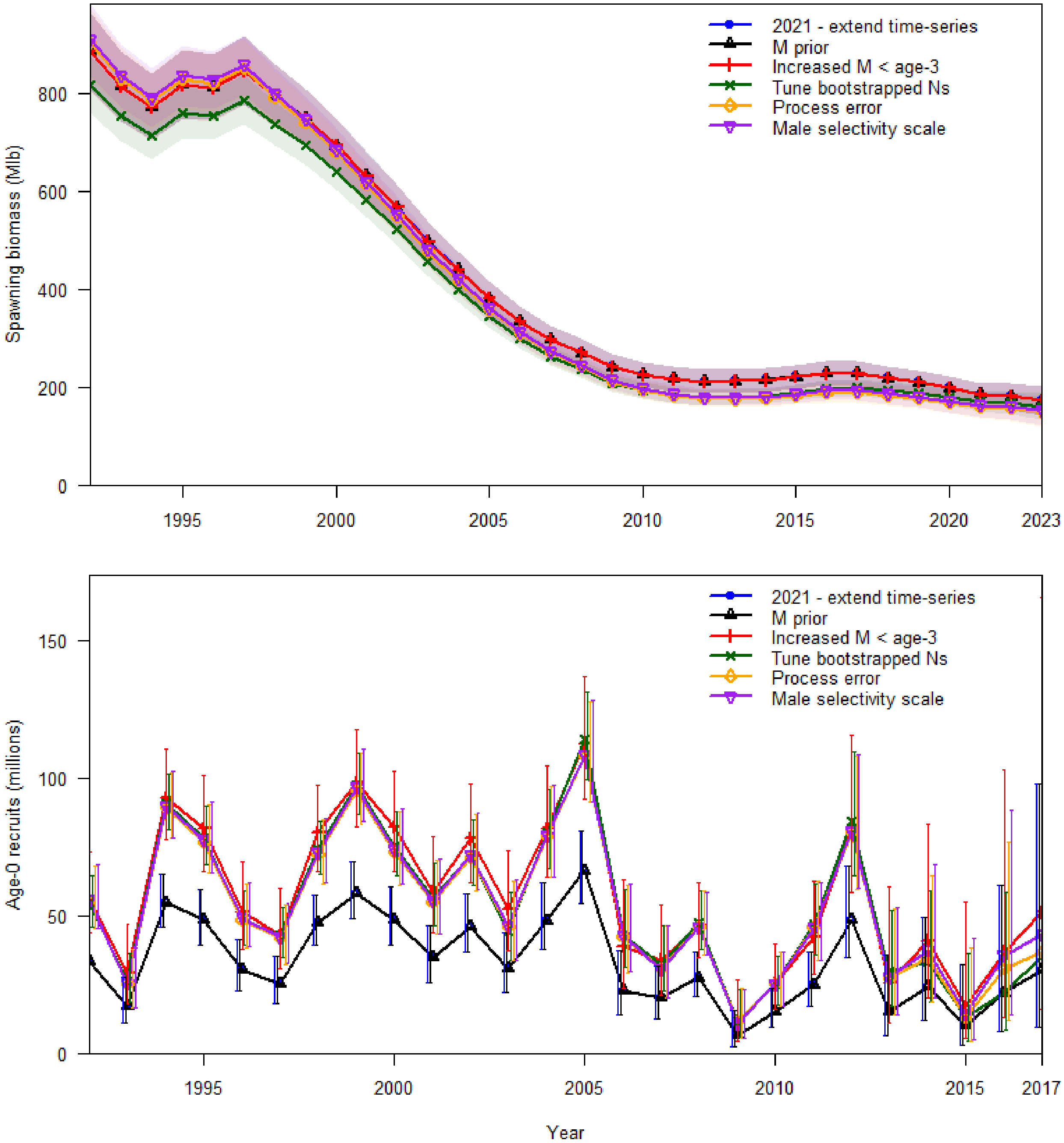


Figure 17. Comparison of estimated biomass (upper panel) and recruitment time series (lower panel) over sequential changes from the 2021 to preliminary 2022 AAF short models.

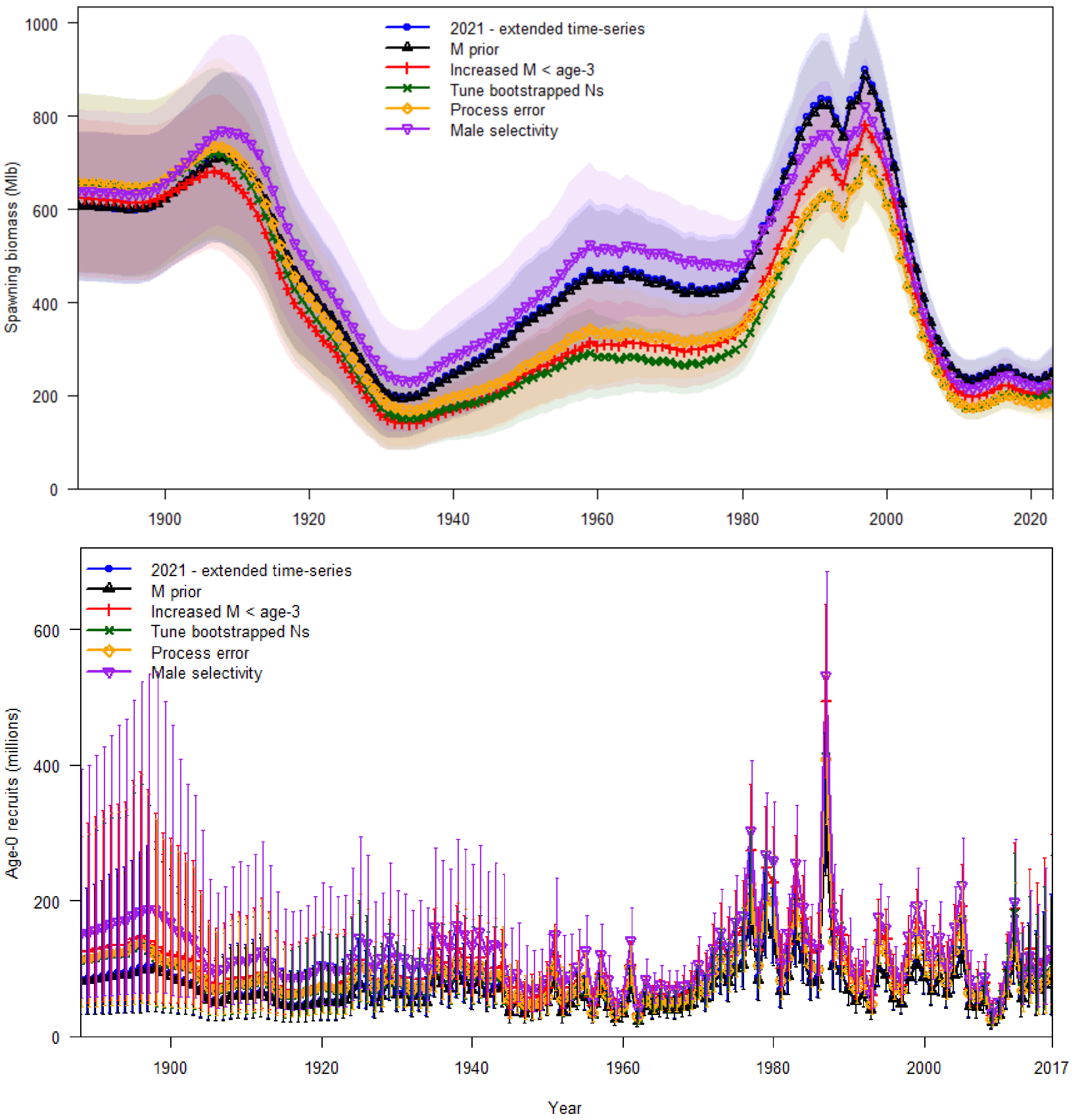


Figure 18. Comparison of estimated biomass (upper panel) and recruitment time series (lower panel) over sequential changes from the 2021 to preliminary 2022 AAF long models.

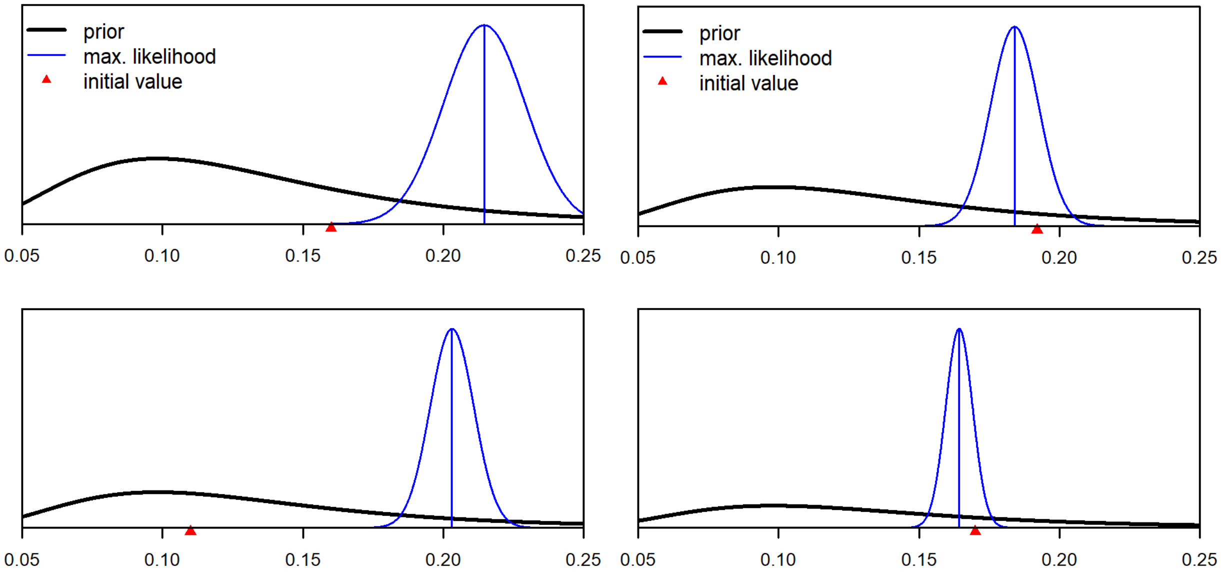


Figure 19. Prior and maximum likelihood estimates for female (upper panels) and male (lower panels) *M* in the coastwide long (left panels) and AAF long (right panels) assessment models.

Commercial fishery selectivity

In the 2019 stock assessment, the AAF short and the two long models did not allow the scale of the male fishery selectivity curve (which mainly determines the sex-ratio of the landings) to be time-varying. At the time of the preliminary assessment there were only 2 years of sex-specific age compositions available. As additional years of data have become available (now 2017-2020, with 2021 anticipated for the full 2022 stock assessment), it is now possible to allow the models to track the year-to-year variability, and more importantly, to disconnect the recent parameter estimates from the historical period. Again here, the results were generally insensitive to this change, with the exception being the most complicated of the models, the AAF long model (Figure 18). While not evident in the bridging analysis, this change created much more stable retrospective patterns than observed in previous assessments (see retrospective section below).

Estimation of female M in the two short models

The final change evaluated in the bridging analysis was the estimation of female *M* in the two short time-series models. For the short coastwide model, all efforts to estimate *M* resulted in the value going to the upper bound. As has been the case in previous assessments, the conclusion was reached that this value was not estimable, even with the informative prior now available. In contrast, the AAF short model produced an estimate of *M* consistent with the two long time-series models and the likelihood surface clearly indicated that the fixed value of 0.15 was much less plausible (0.21, see likelihood profile section below). This step in the bridging analysis is plotted separately along with the previous and initial step so that the results can be more clearly compared (Figure 20).

The choice to fix or estimate female *M* is an important one, which has clear implications for the scale of the estimated spawning biomass. Previous short time-series assessments, back at least to 2006, either assumed that female *M* was not estimable or did not find a clear minimum within the range of values considered plausible. There is no clear basis for the historically assumed value of 0.15, but the choice to fix female *M* has led to models with very tight uncertainty intervals, in contrast to the much broader intervals estimated here (Figure 20).

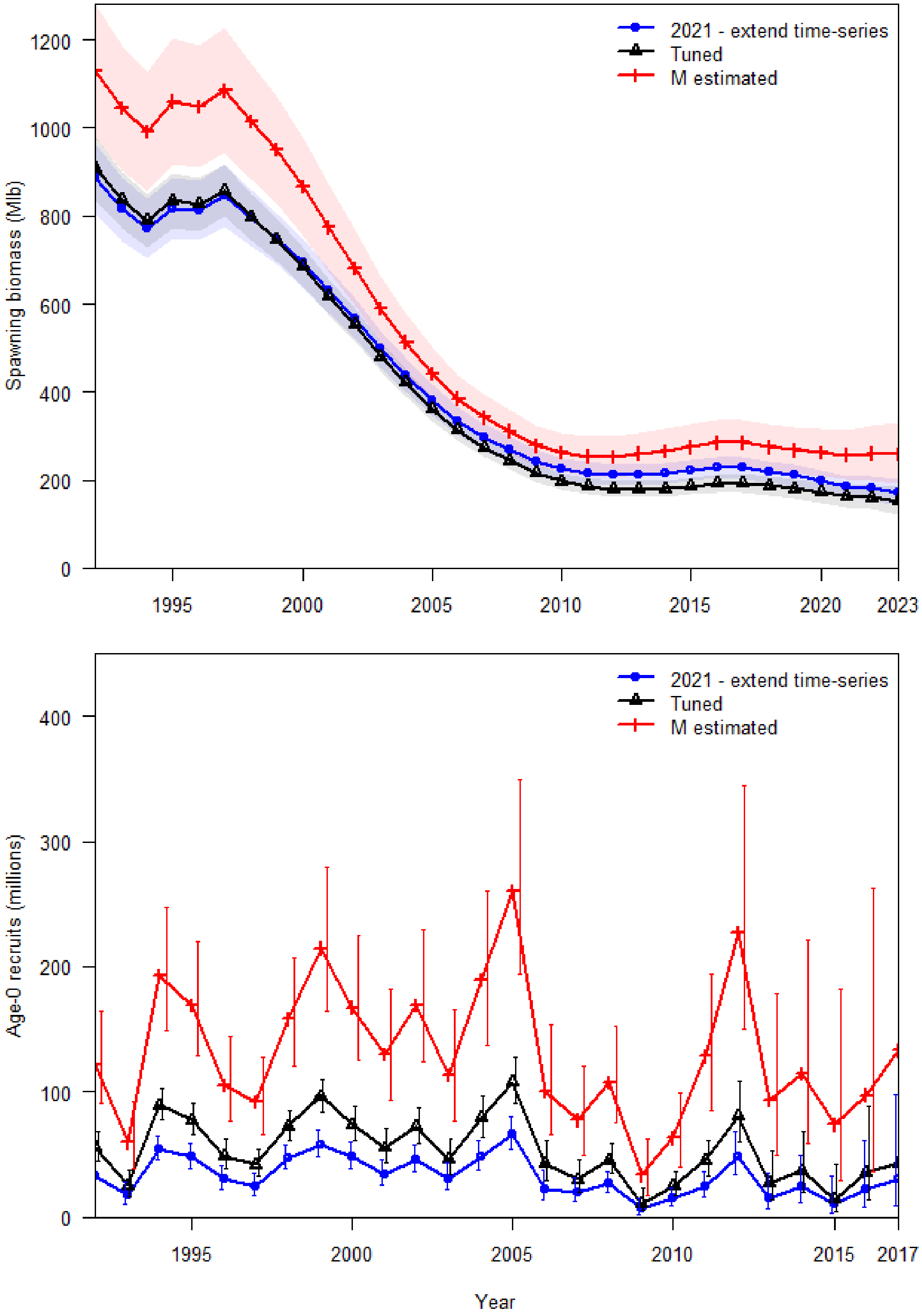


Figure 20. Comparison of estimated biomass (upper panel) and recruitment time series (lower panel) for AAF short models with and without female *M* estimated.

Convergence criteria

Standard tools for monitoring convergence criteria include assessing the maximum gradient component, sensitivity to alternative phasing and initial values, use of overdispersed starting points or 'jitter analyses', as well as likelihood profiles, and Bayesian integration.

For this preliminary 2022 assessment, all individual models all had a maximum gradient component < 0.004. A series of preliminary and intermediate runs did not indicate any signs that the estimates reported here represented local minima for all but the AAF long model, nor did the models have difficulty converging and producing a positive definite Hessian matrix under the range of alternative and sensitivity analyses (some presented in this document, but many used only for development).

Convergence was tested specifically through a 'jitter' analysis perturbing all parameter values simultaneously by 1% of the range between upper and lower bounds and repeating minimization. Initial testing revealed that the coastwide long model recovered the MLE 100% of the time. Similarly, the coastwide short model recovered the MLE 98% of the time and failed to converge to a solution 2% of the time. Being more complex, convergence success was lower for the AAF short model, recovering the MLE 68% of the time, failing to converge 21% and stopping short of the actual MLE 11%. The AAF long model, with considerably more process-error parameters than the others did show a greater sensitivity to all sensitivity and bridging analyses. Further, this model did occasionally get stuck at an alternate minimum that was 1% different in spawning biomass and 1.16 negative log likelihood units worse than the true minimum. The AAF long model required starting values much closer to the true MLE for a wide range of runs, and still converged to the MLE 44% of the jittered runs, 21% stopping short and 35% failing to converge. This indicates that, at least for the current configuration, use of good starting values and jitter analyses is most important for the long AAF model.

Wherever parameters were hitting bounds either the bounds were adjusted (if biologically plausible) or the parameters were fixed. For example, the descending limb of the 4B commercial fishery in the AAF models was estimated to be at the bound of 1.0 (as has been the case for all recent assessments), and so was fixed at this value. This approach reduces the likelihood that variances calculations will be (undesirably) effected by parameters stuck to bounds but does require periodic revisitation to ensure that the signal for parameters hitting bounds remains, and that fixing those parameters does not have an appreciable effect on the maximum likelihood solution.

Individual model diagnostics and results

This section provides more detail on the specific diagnostics and results of each of the four assessment models. It is not intended to provide the fit and residuals to every data component, but to summarize the basic performance of the model and specifically highlight areas of potential deficiency. Figures showing comprehensive diagnostics and results and the full report files, as output directly from stock synthesis, are provided electronically as described in Appendix A. Each model section finishes with a brief summary of the relative strengths and weaknesses of that model.

Coastwide short

Predictions of both the fishery and survey indices of abundance fit the observed data very well in the coastwide short model (Figure 21). In the 2018 assessment, a small amount of process error was allowed on fishery catchability. Since 2019, the iterative tuning of the annual catchability deviations suggested that process error was no longer needed. The predicted aggregate age distributions also matched the observed distributions well, for both the fishery and survey indicating that the selectivity parameterization was generally capturing differences in both the age-structure and the sex-ratio (Figure 22).

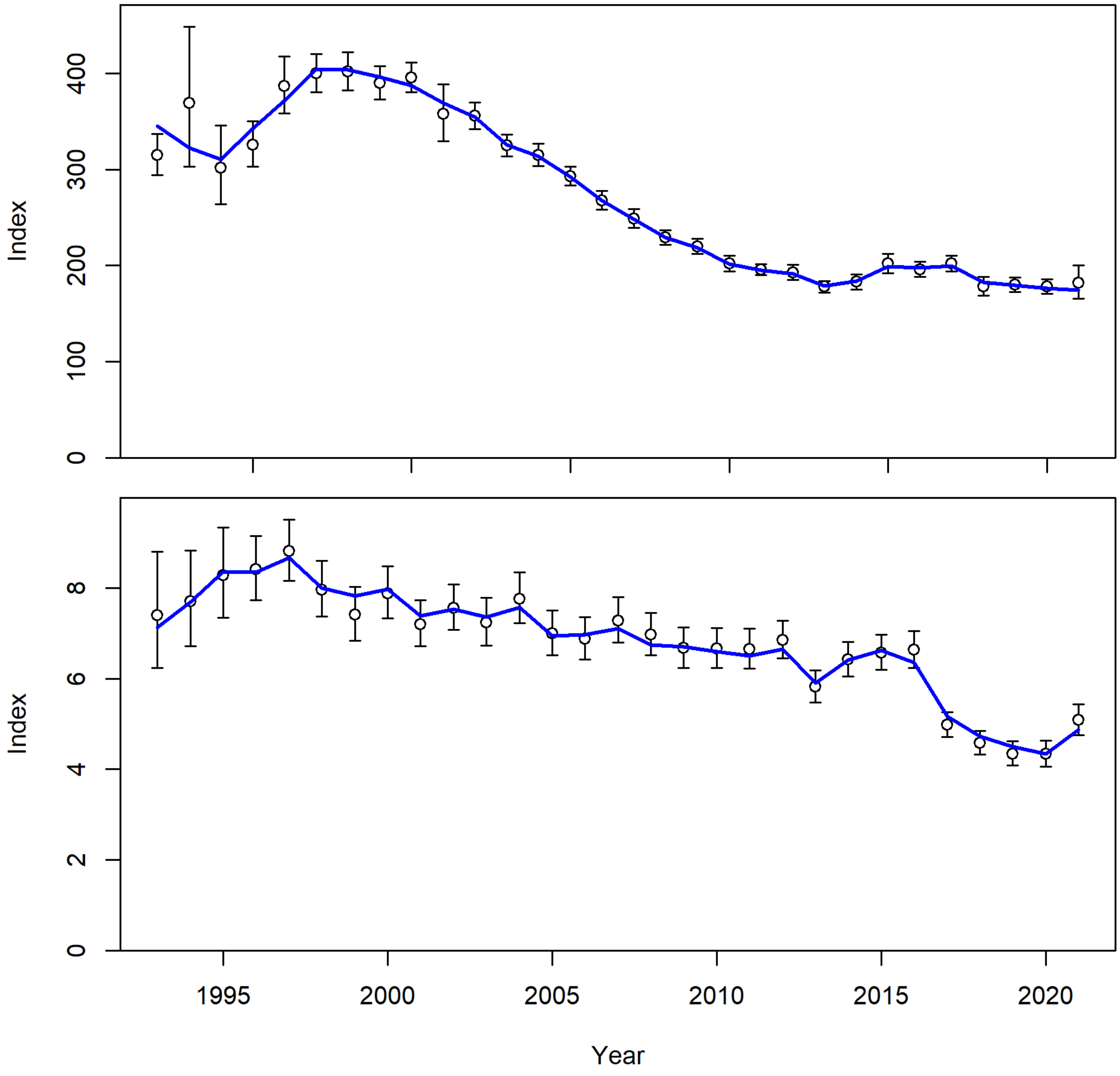


Figure 21. Fit to fishery (upper panel) and FISS (lower panel) indices of abundance in the coastwide short model; note that the scale of the y-axes differ.

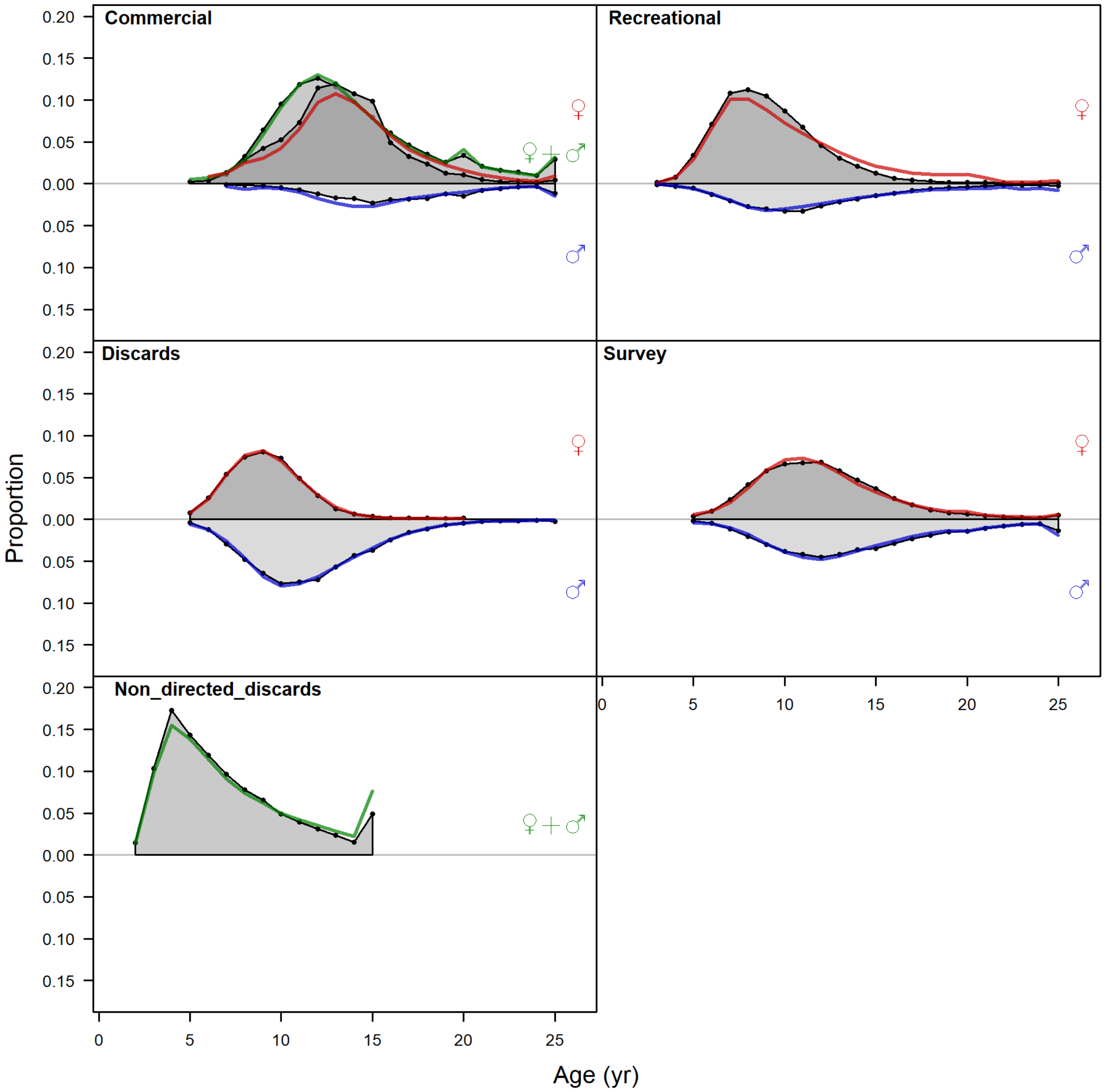


Figure 22. Aggregate fit to all age data by model fleet in the coastwide short model; sex-specific distributions for the commercial fishery represent only 2017-2020 and are plotted on top of sexes-aggregated distributions spanning 1992-2016 + 2021.

The coastwide short model tuning resulted in a higher weight on the coastwide FISS ages than for the commercial fishery age data (Table 16). The discard, non-directed discard and recreational age data were all heavily down-weighted (as described above) and so input sample sizes were not iterated to larger values, despite fits to the data that implied a higher weight. Fit to the annual FISS age compositions were generally good (Figure 23), although some patterning was visible in the standardized residuals (Figure 24). Specifically, there was a clear pattern of negative residuals in the plus group for male halibut; however, this was almost imperceptible in the fits themselves due to the very small observed and predicted values in this age bin. The fits to the annual fishery data were also acceptable (Figure 25-25). Similarly, the implied fit to the sex ratio information for the commercial fishery was somewhat more variable (Figure 27) than that for the FISS (Figure 28). Additional diagnostics and diagnostic figures (such as fits to the down-weighted annual compositions for the discard, bycatch, and recreational fleets) are included in the background materials.

Table 16. Post-iteration sample size diagnostics for age-composition data by model and fleet. Average iterated input denotes the value used for model runs reported here, after iterating the bootstrapped starting points.

	Average iterated input	Harmonic mean effective	Francis weight effective	Maximum Pearson residual
Coastwide short				
Fishery	62	294	62	2.45
Discards[1]	13	270	49	0.98
Non-directed discards[1]	5	47	39	2.25
Recreational[1]	5	114	27	0.88
FISS	242	668	242	2.06
Coastwide long				
Fishery	112	289	122	4.09
Discards[1]	6	210	90	0.78
Non-directed discards[1]	3	37	7	1.33
Recreational[1]	3	145	31	0.51
FISS	82	194	83	2.88
AAF short				
Region 2 fishery	723	676	1,078	4.47
Region 3 fishery	808	699	951	3.85
Region 4 fishery	23	78	36	3.54
Region 4B fishery[2]	36	138	81	1.82
Discards[1]	13	219	73	1.21
Non-directed discards[1]	5	58	22	1.12
Recreational[1]	5	143	20	0.85
Region 2 FISS	7	86	7	1.04
Region 3 FISS	18	262	18	1.25
Region 4 FISS	66	181	63	3.95
Region 4B FISS[2]	41	185	50	1.83
AAF long				
Region 2 fishery	322	304	651	4.31
Region 3 fishery	266	309	544	3.78
Region 4 fishery	18	60	28	4.36
Region 4B fishery[2]	37	129	80	1.90
Discards[1]	6	189	84	1.56
Non-directed discards[1]	3	43	8	1.12
Recreational[1]	8	151	23	0.91
Region 2 FISS	7	78	8	1.39
Region 3 FISS	12	101	13	1.26
Region 4 FISS	72	182	68	3.53
Region 4B FISS[2]	41	185	45	1.93

[1] Inputs down-weighted, and not iteratively reweighted – see text.

[2] Iterated sample size equal to maximum (bootstrapped input).

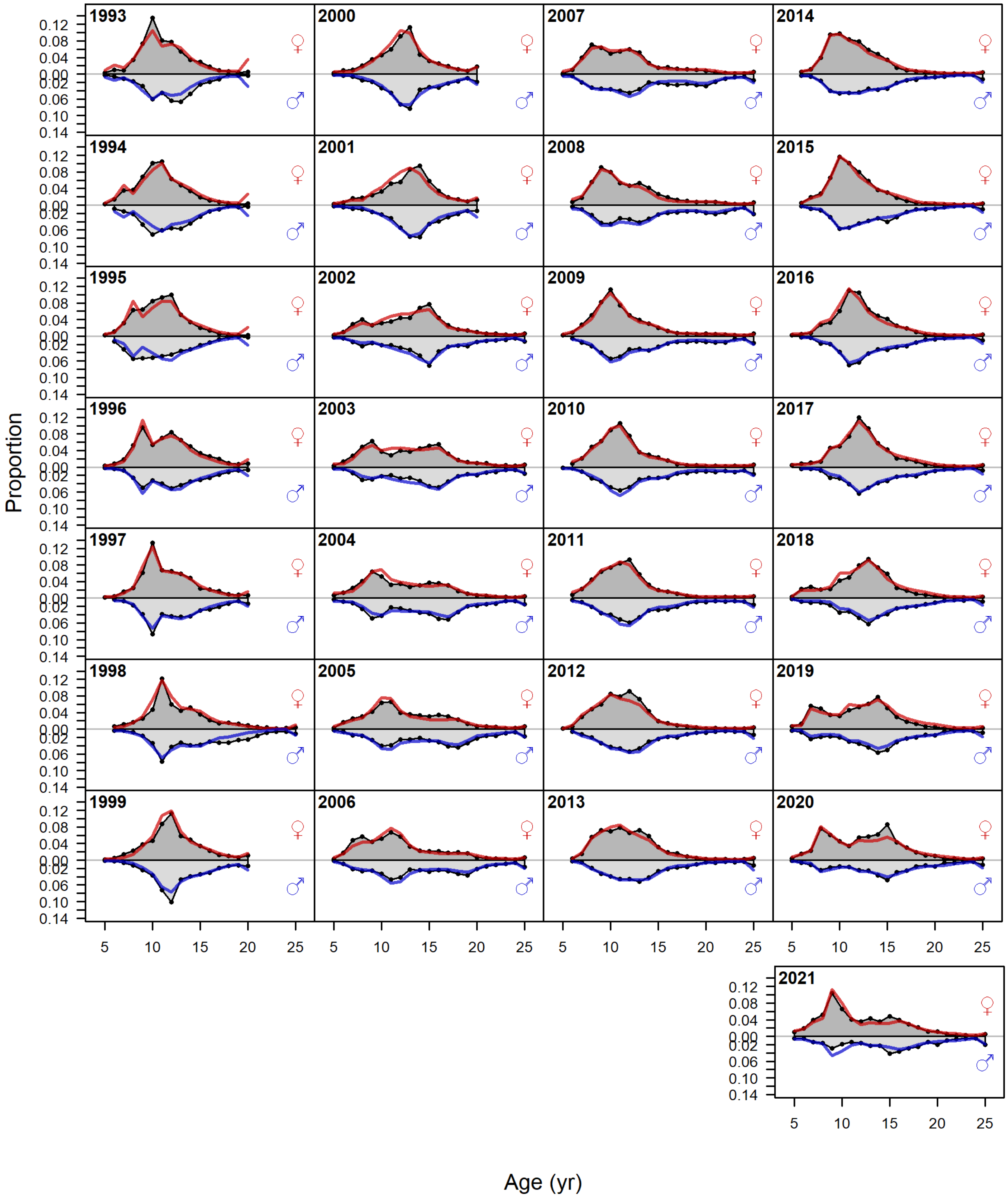


Figure 23. Fit to annual age data from the FISS survey in the coastwide short model.

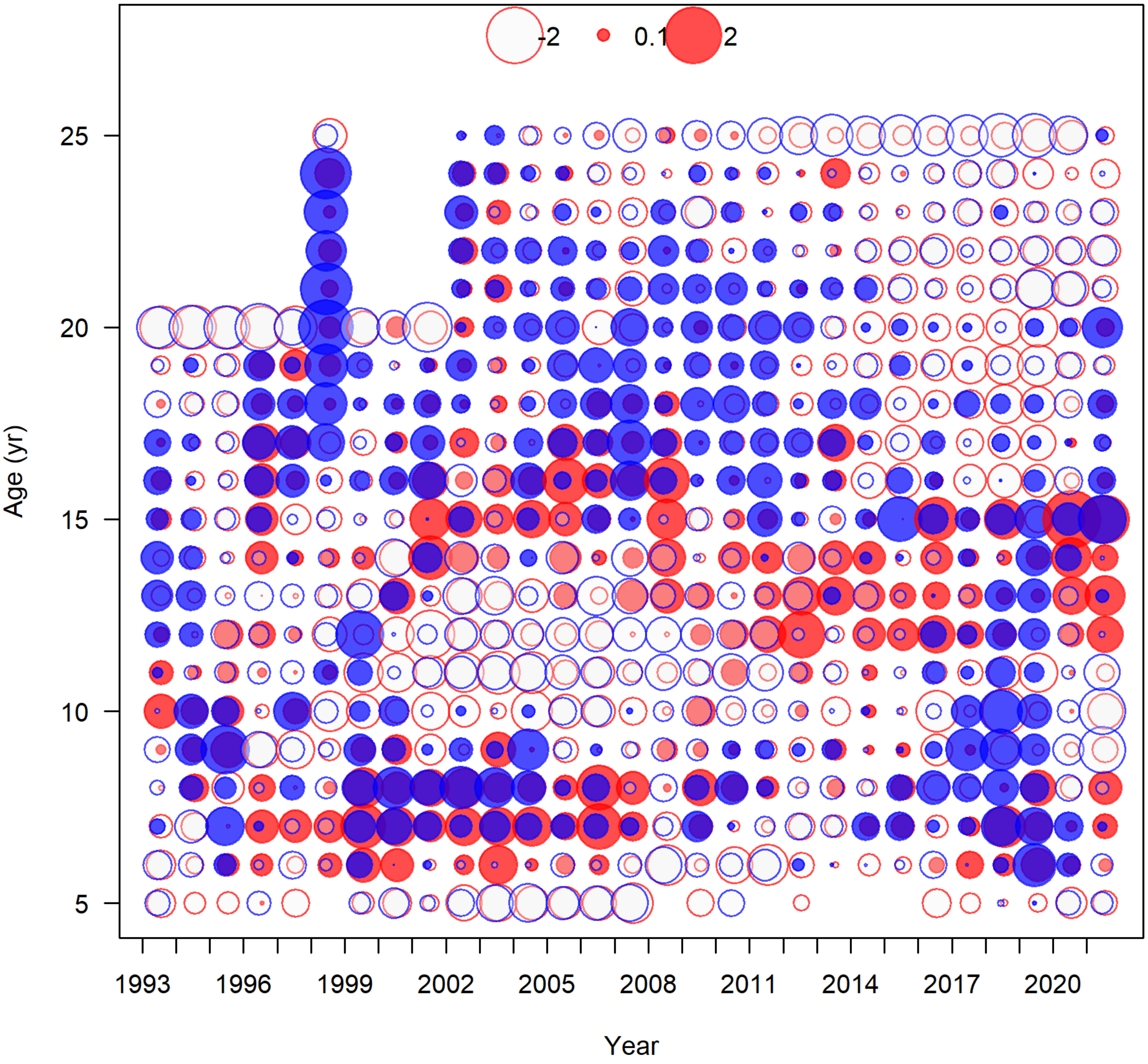


Figure 24. Pearson residuals for fit to annual age data from the FISS survey in the coastwide short model; red circles denote female residuals, and blue circles denote male residuals.

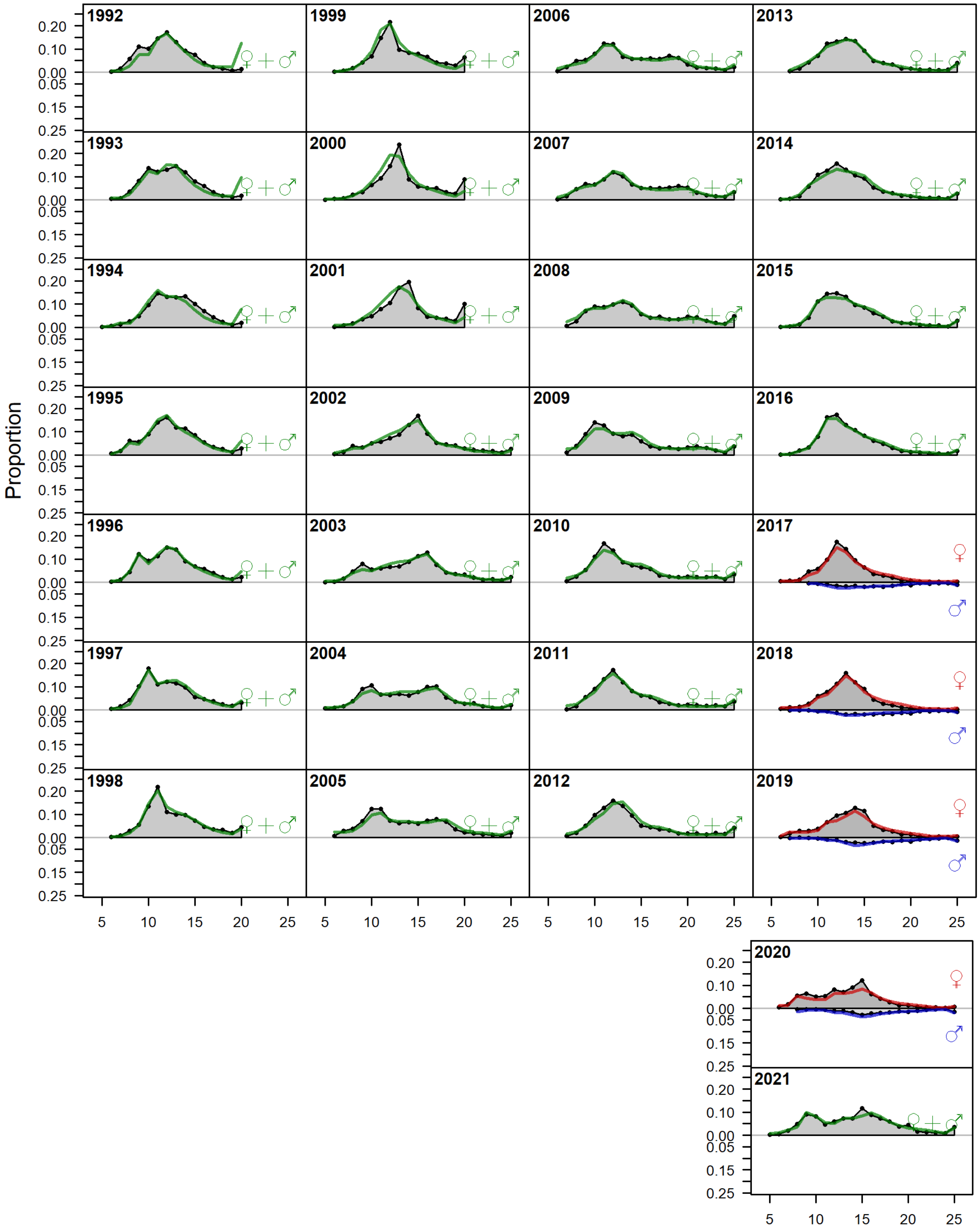


Figure 25. Fit to annual age data from the commercial fishery landings in the coastwide short model.

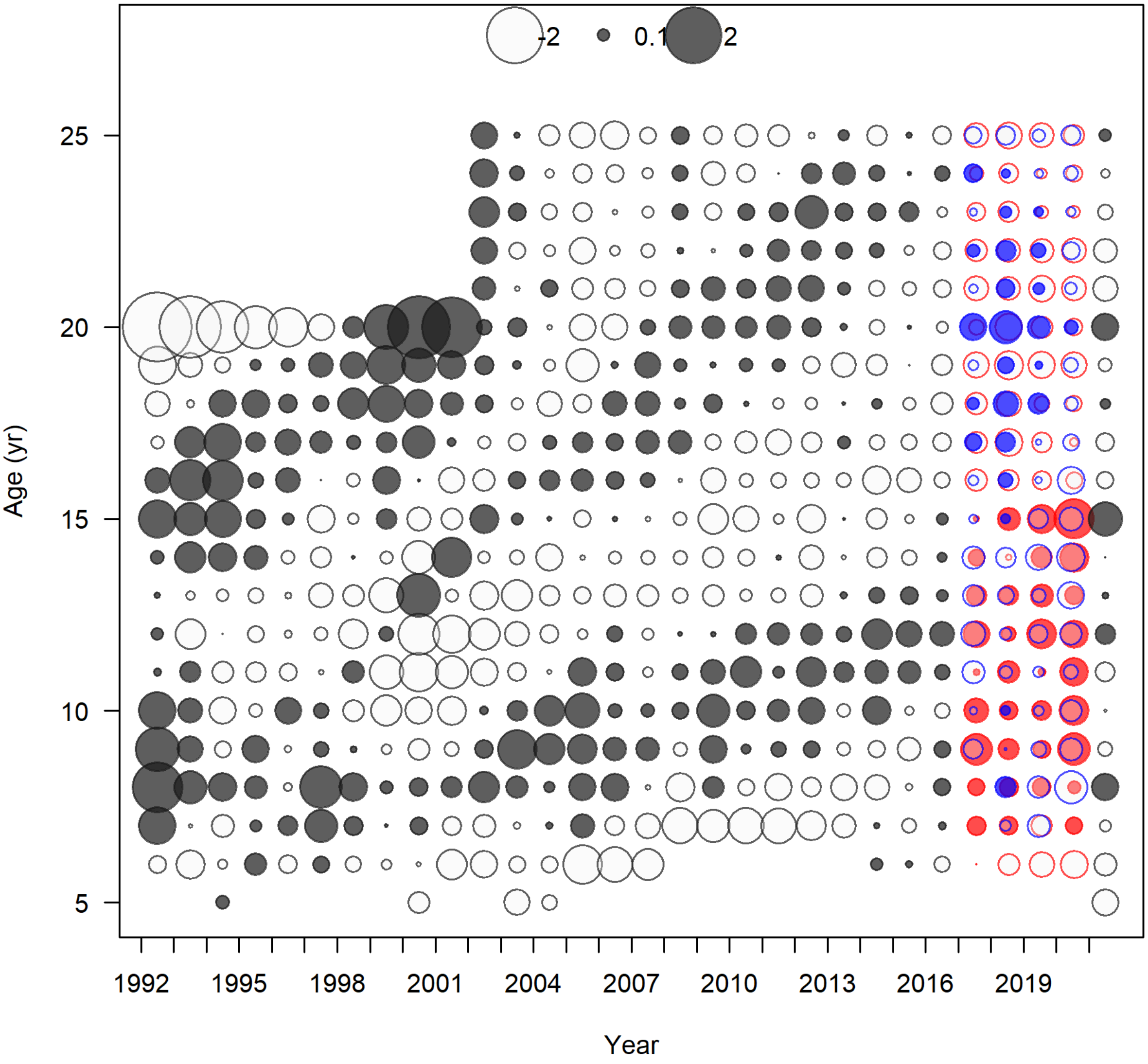


Figure 26. Pearson residuals for the fit to annual age data from the commercial fishery landings in the coastwide short model; grey circles denote unsexed residuals, red circles denote female residuals, and blue circles denote male residuals.

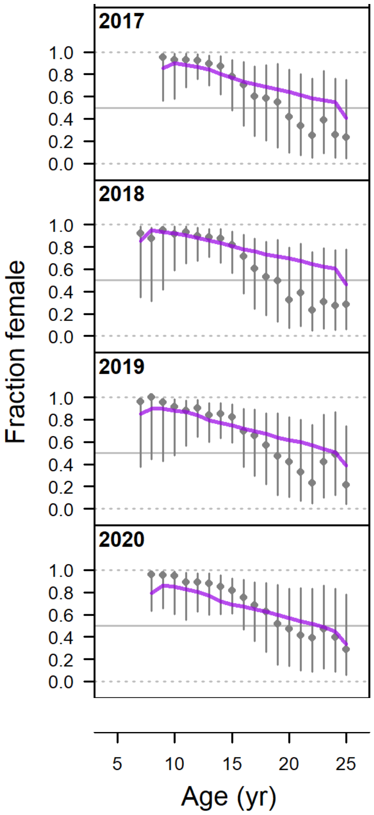


Figure 27. Observed and predicted sex-ratio in the commercial fishery landings from the coastwide short model.

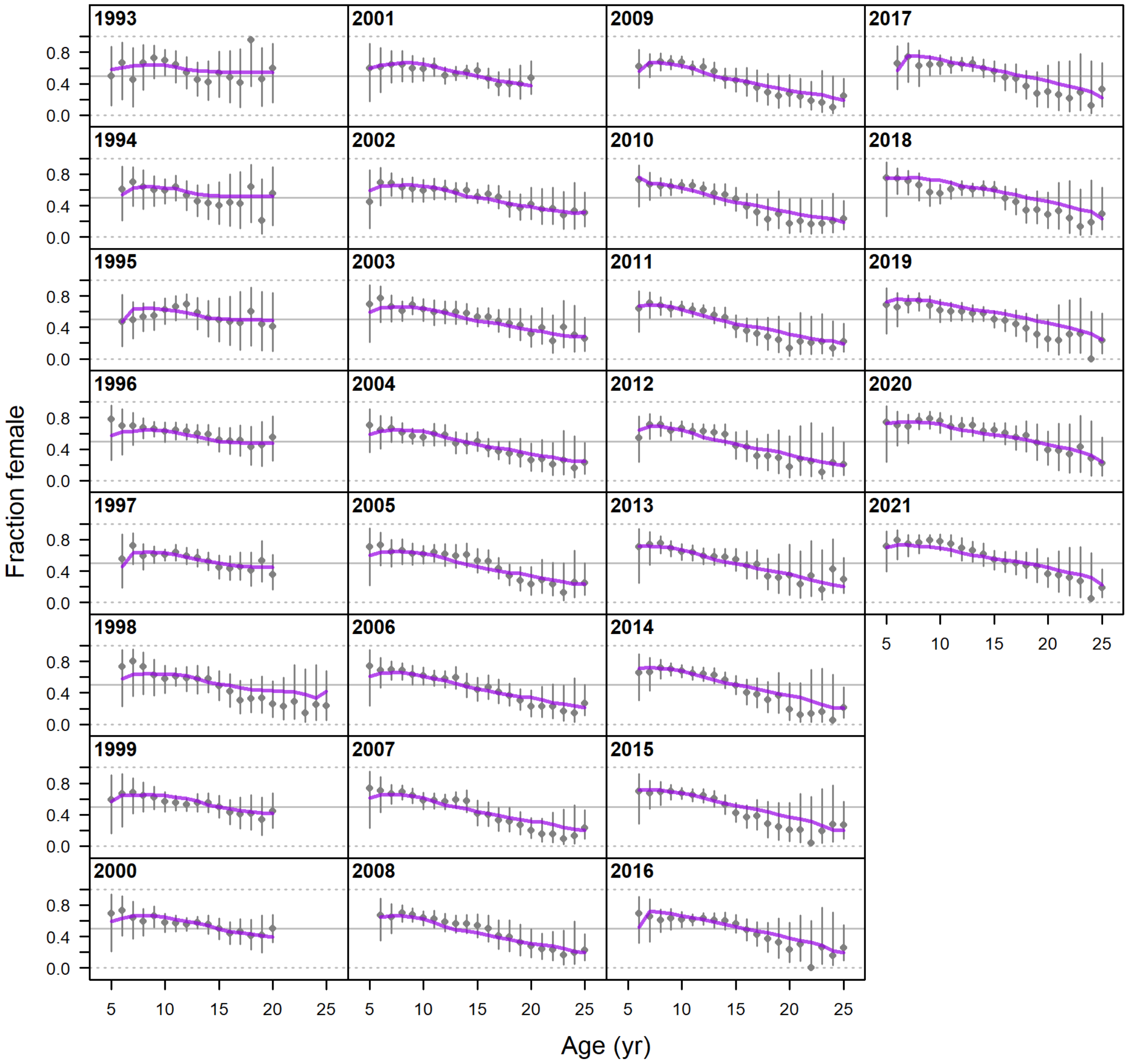


Figure 28. Observed and predicted sex-ratio in the FISS from the coastwide short model.

Neither the FISS nor the fishery selectivity was estimated to have a highly variable ascending limb over the short time-series (Figure 29). The estimated fishery selectivity showed a trend toward increasing selection of males in the middle of the time-series, more pronounced than that estimated for the FISS (Figure 30), perhaps a function of the catch distribution shifting toward the Eastern side of the stock where fast-growing males are much more common, as well as the decline in the strong cohorts from the 1980s which produced an abundance of older females. For the discard fleet, estimated selectivity included fewer males than females (Figure 31). Estimated selectivity for the non-directed discards fleet showed a peak at ages 4-5 and a slightly domed relationship. Recreational/subsistence selectivity was shifted to the left of the commercial fishery discards (and therefore the FISS).

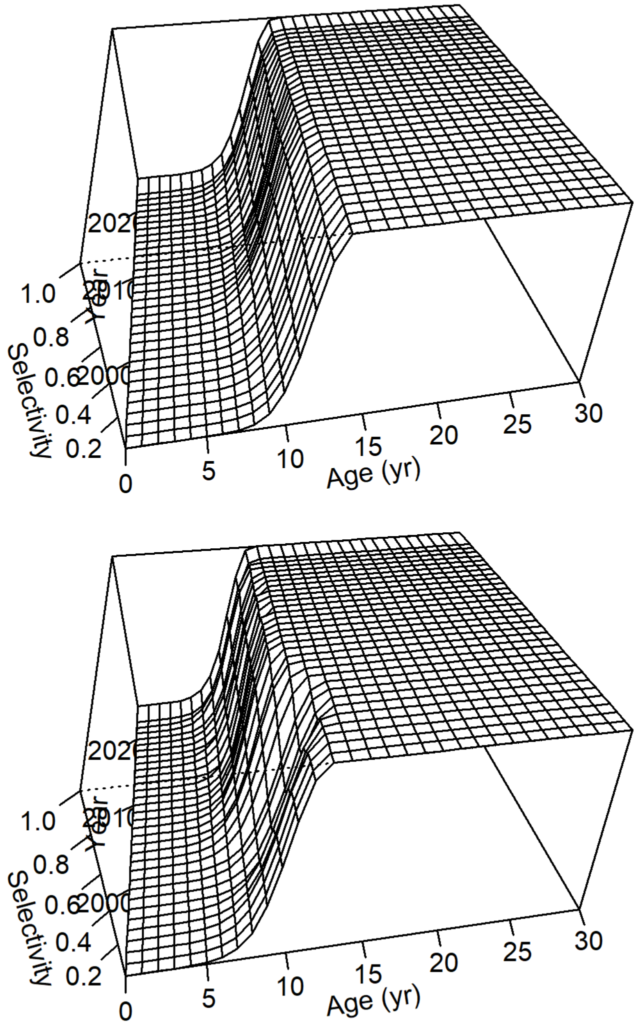


Figure 29. Estimated time-varying female selectivity curves for the commercial fishery landings (upper panel) and the FISS (lower panel).

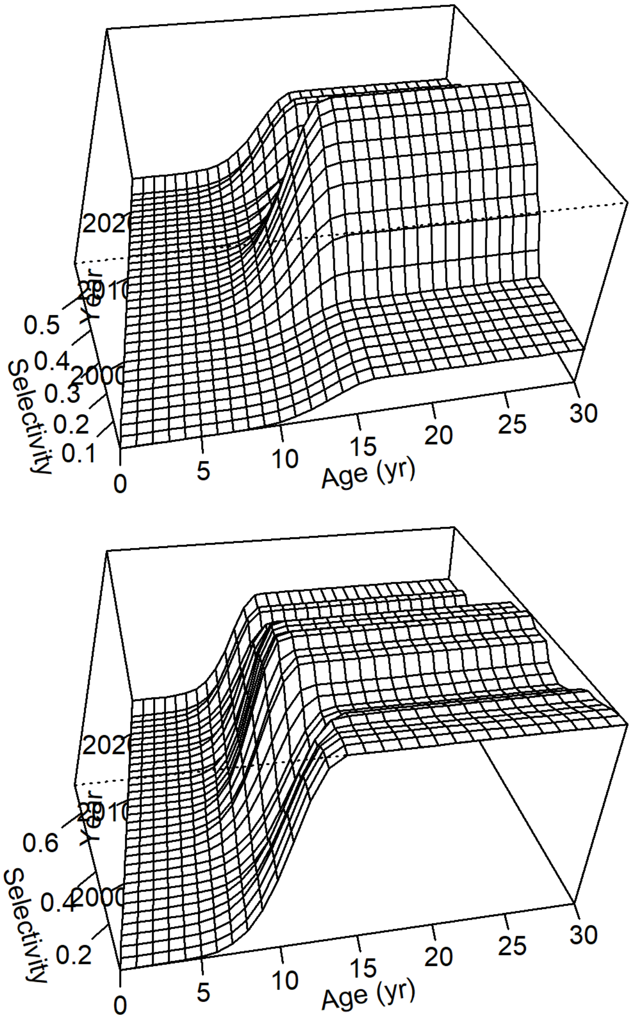


Figure 30. Estimated time-varying male selectivity curves for the commercial fishery landings (upper panel) and the FISS (lower panel).

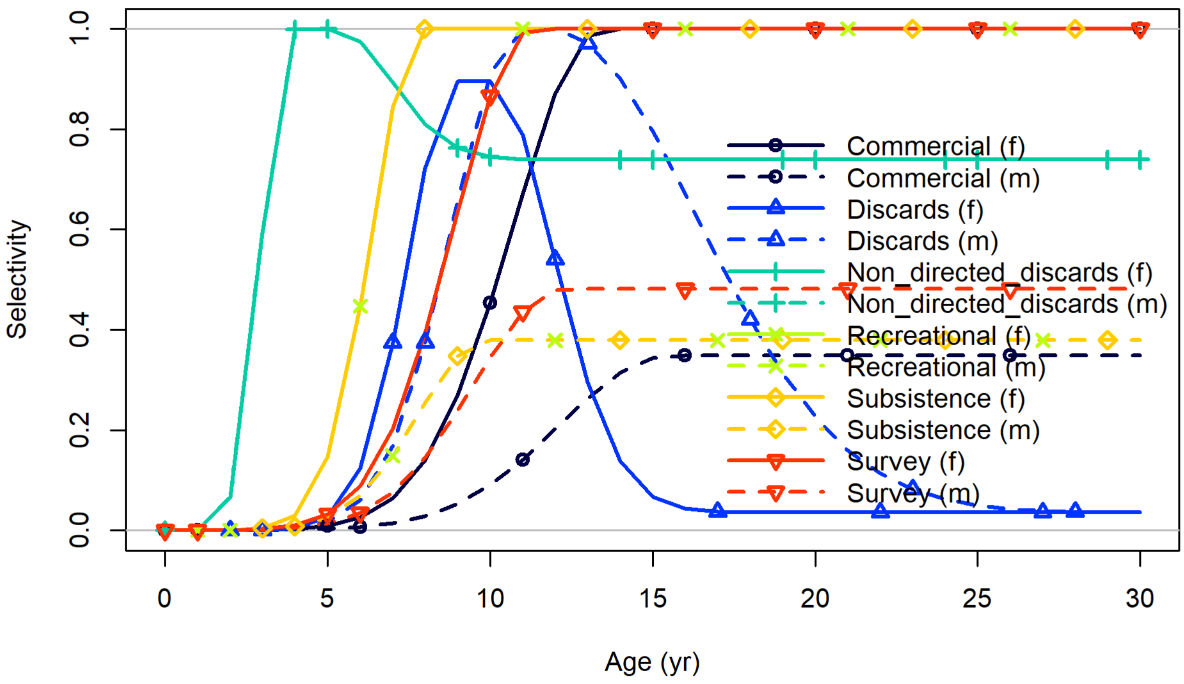


Figure 31. Estimated ending year selectivity curves by sex for the commercial fishery, discard, non-directed discard, recreational and FISS fleets in the coastwide short model.

Male *M* was estimated to be slightly lower (0.149) than the fixed value assumed for females of 0.15 (Table 17); this represented a slight increase from the value estimated in the 2019 and earlier assessments. The large negative estimated initial recruitment offset is consistent with the start year occurring after a very long time-series of fishing. The lower *M* fixed in the coastwide short model corresponded to much lower recruitment and female spawning biomass estimates (Table 17) than the other three models, as has been the case for all recent assessments.

Summary of strengths and weaknesses for the coastwide short model:

Strengths:

- Lowest technical overhead (complexity) of the four models in the ensemble
- Fit the fishery and FISS indices very well
- Fit the survey age data (males and females) relatively well
- Parameter estimates are derived from the most recent time period
- Internally consistent data weighting

Weaknesses:

- Basis for fixed female *M* is unclear
- Does not include uncertainty in female *M* (see likelihood profile evaluation below)
- Does not include extensive historical data
- May lose Region-specific trends and biological patterns due to aggregation

- Does not use environmental information to inform recruitment
- Commercial fishery age data is not heavily weighted and there are therefore residual patterns despite allowing for process error in selectivity

Table 17. Select parameter estimates (maximum likelihood value and approximate 95% confidence interval) and important recent population estimates by model and Biological Region (where applicable).

	Model			
	Coastwide Short	Coastwide Long	AAF Short	AAF Long
Biological				
Female *M*	0.150 (*Fixed*)	0.215 (0.186-0.243)	0.211 (0.195-0.227)	0.184 (0.167-0.200)
Male *M*	0.149 (0.138-0.159)	0.203 (0.188-0.218)	0.177 (0.167-0.187)	0.164 (0.154-0.173)
Log(R_0)	11.375 (11.167-11.582)	11.857 (11.546-12.168)	12.347 (12.115-12.579)	11.545 (11.262-11.829)
Initial log(R_0) offset	-1.469 (-1.685--1.253)	NA	-0.368 (-0.596-0.140)	NA
Environmental Link (β)	NA	0.372 (0.144-0.600)	NA	0.349 (0.129-0.569)
Survey Log(q) Δ1984 (transition to circle hooks)	NA	0.945 (0.592-1.299)	NA	R2: 1.222 (0.844-1.600) R3: 1.822 (1.553-2.092)
Fishery Log(q) Δ1984	NA	0.718 (0.541-0.895)	NA	R2: 0.586 (0.402-0.769) R3: 0.920 (0.724-1.115) R4: 0.858 (0.663-1.053) R4B: 0.529 (0.347-0.712)
2012 Recruitment (Millions)	85 (58-112)	283 (127-439)	278 (163-393)	195 (119-270)
2022 SB (Million lb)	150 (126-173)	202 (155-250)	259 (199-320)	218 (178-260)

Coastwide long

Both the fishery and FISS indices were fit well (Figure 32), with breaks in catchability to accommodate the change from “J” to circle hooks which were very large in both series (Table 17). In aggregate, the predicted age compositions matched the observed data well (Figure 33); however, there were notable differences among years within the time-series. Fits to the FISS were quite poor in the early portion of the time series (Figure 34), improving where the data became more spatially comprehensive in the mid-1990s, and quite good in the most recent years (Figure 35). Fishery data fit reasonably well for the entire time-series (Figure 36-36), with patterns in the residuals corresponding to relatively small differences with observed distributions.

The small contribution of males to the fishery landed catch is clear from the four years that have sex-specific information Figure 37. Harmonic mean effective sample sizes were much larger than adjusted inputs when Francis weights were close to 1.0 (Table 16).

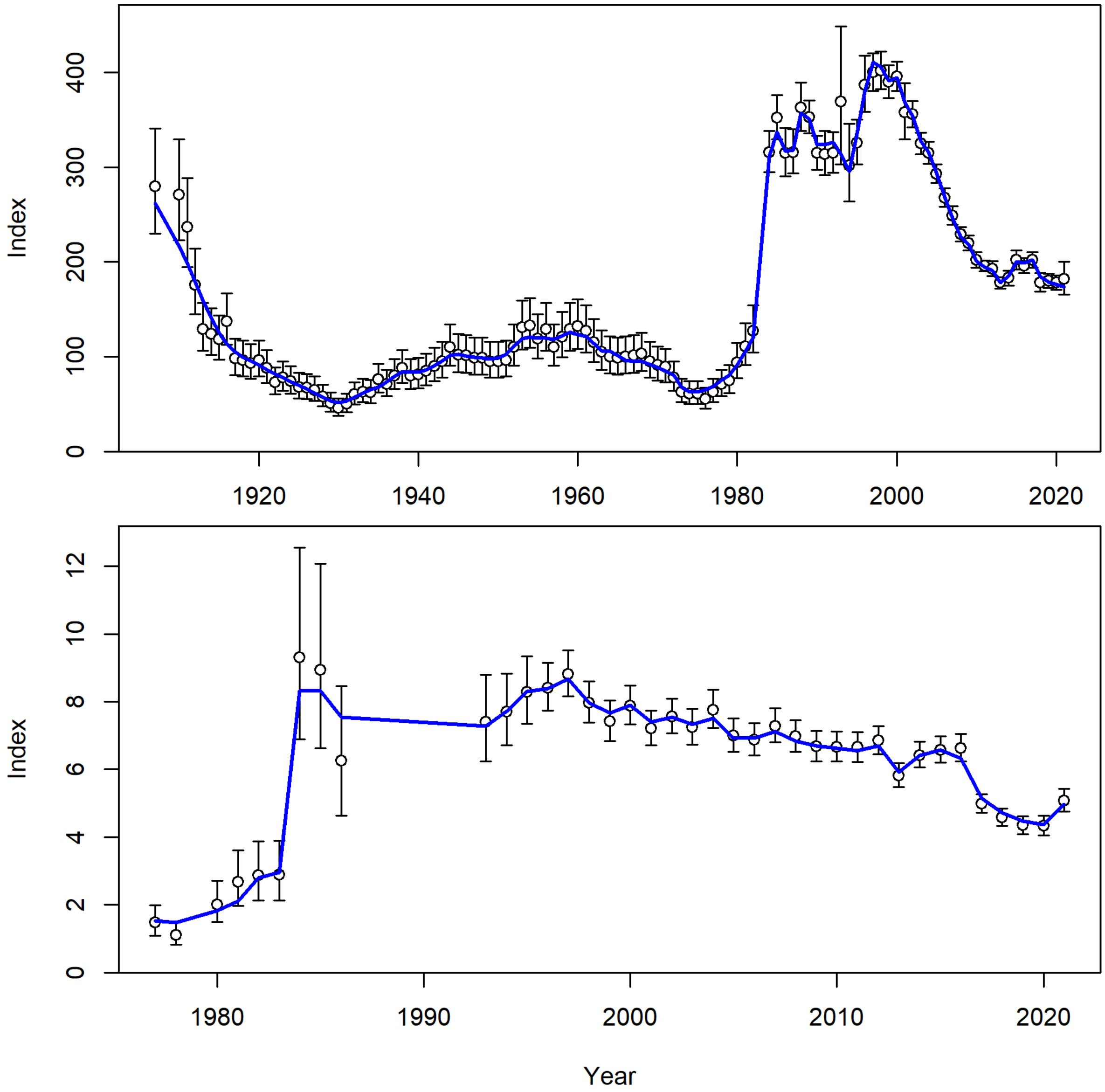


Figure 32. Fit to fishery (upper panel) and FISS (lower panel) indices in the coastwide long model.

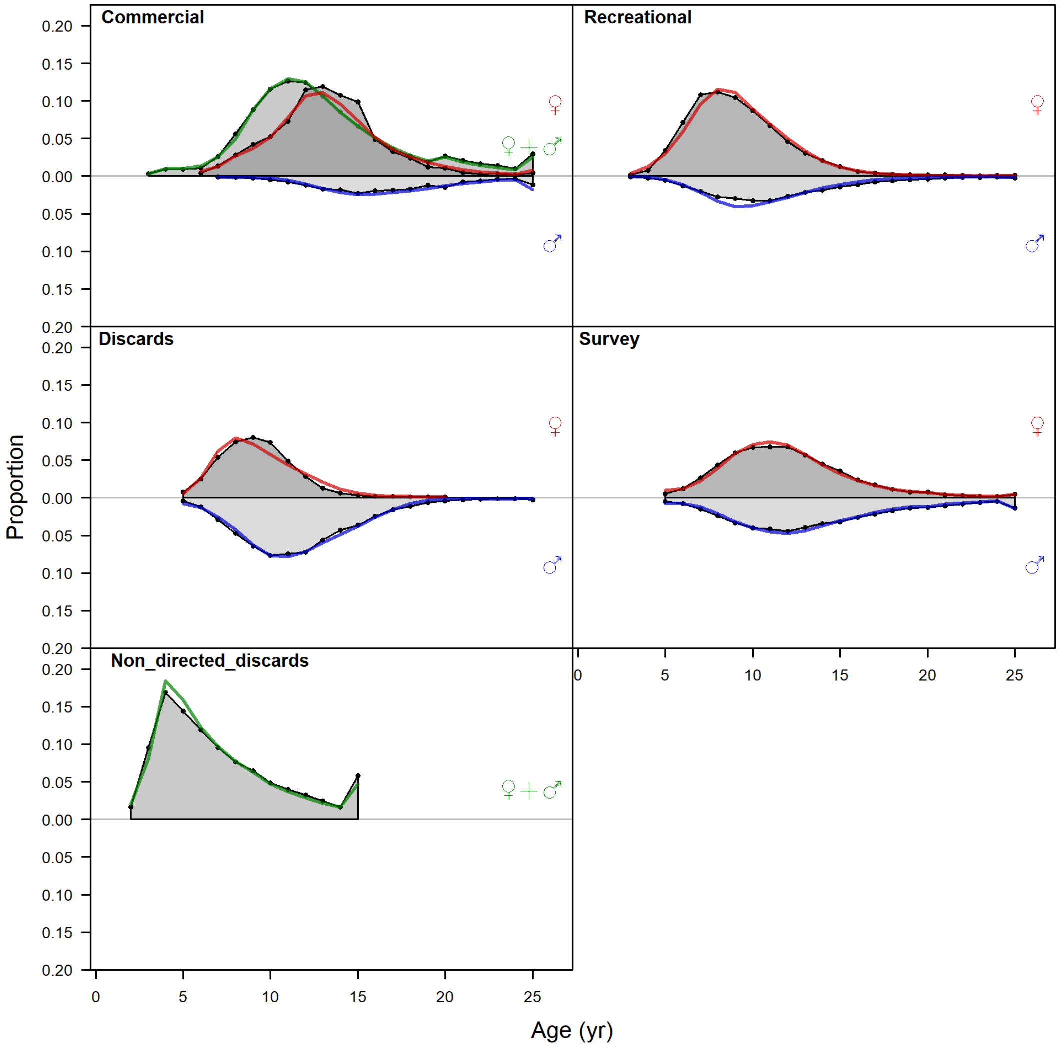


Figure 33. Aggregate fit to all age data by model fleet in the coastwide long model..

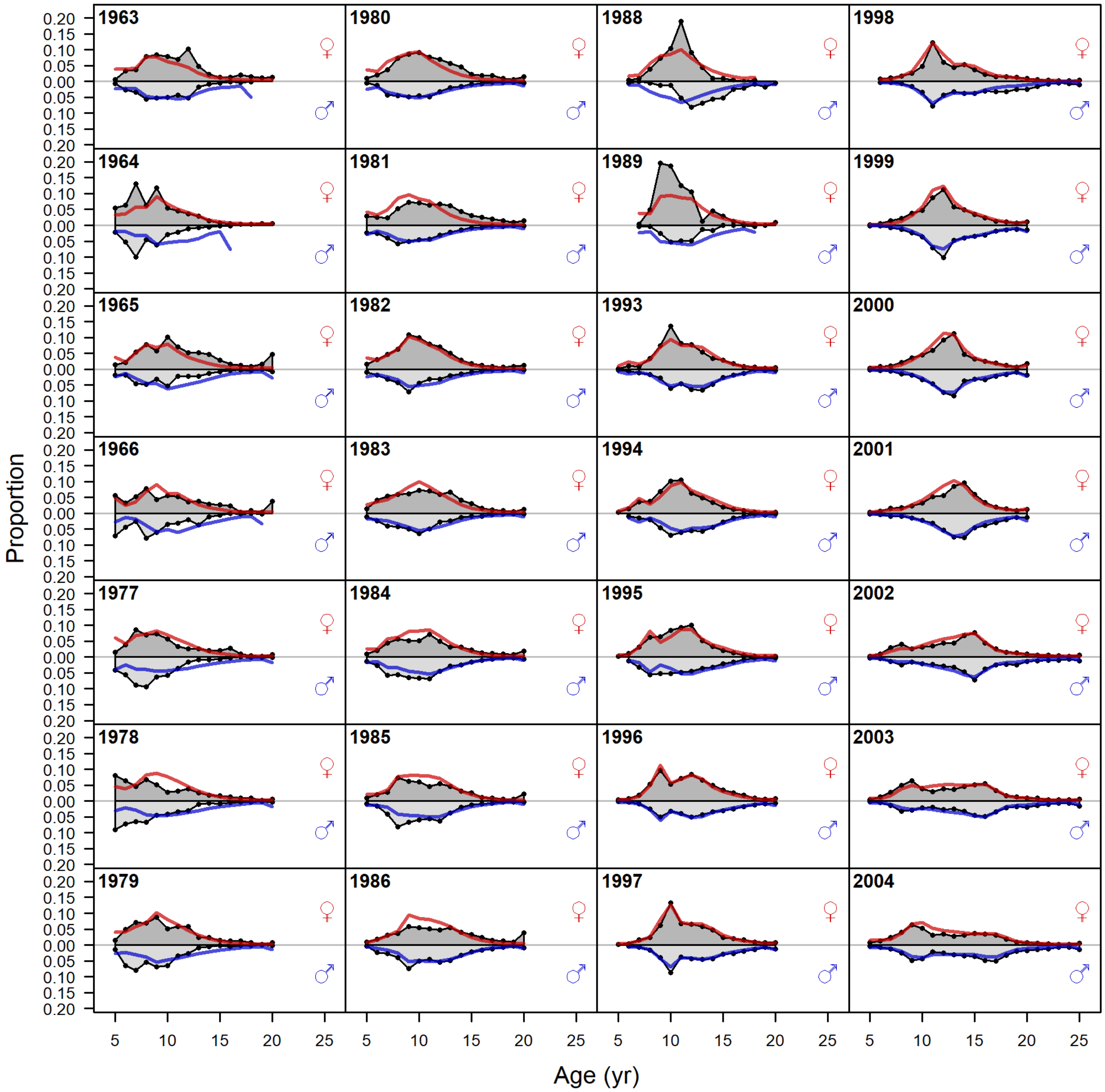


Figure 34. Fit to early years of FISS age data in the coastwide long model.

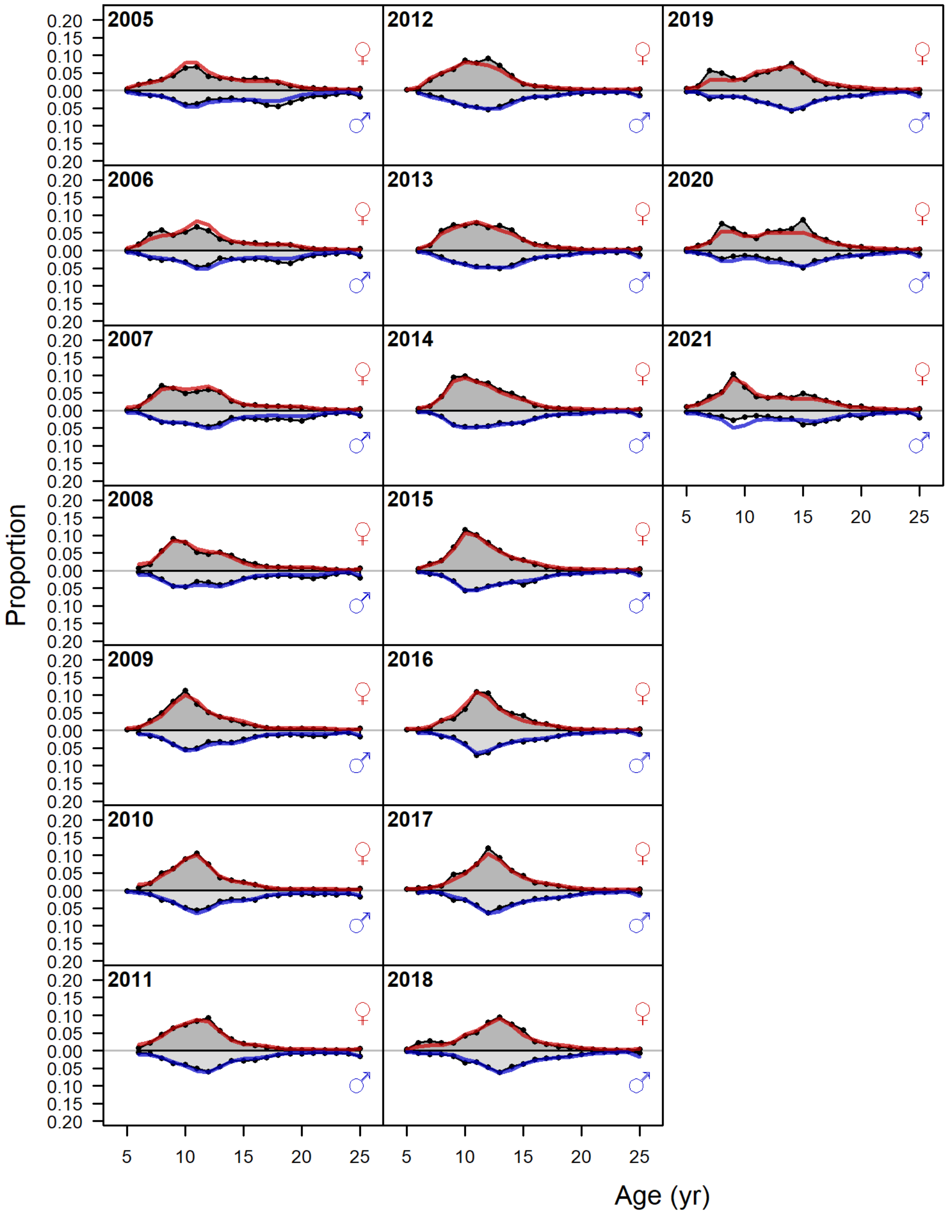


Figure 35. Fit to later years of FISS age data in the coastwide long model.

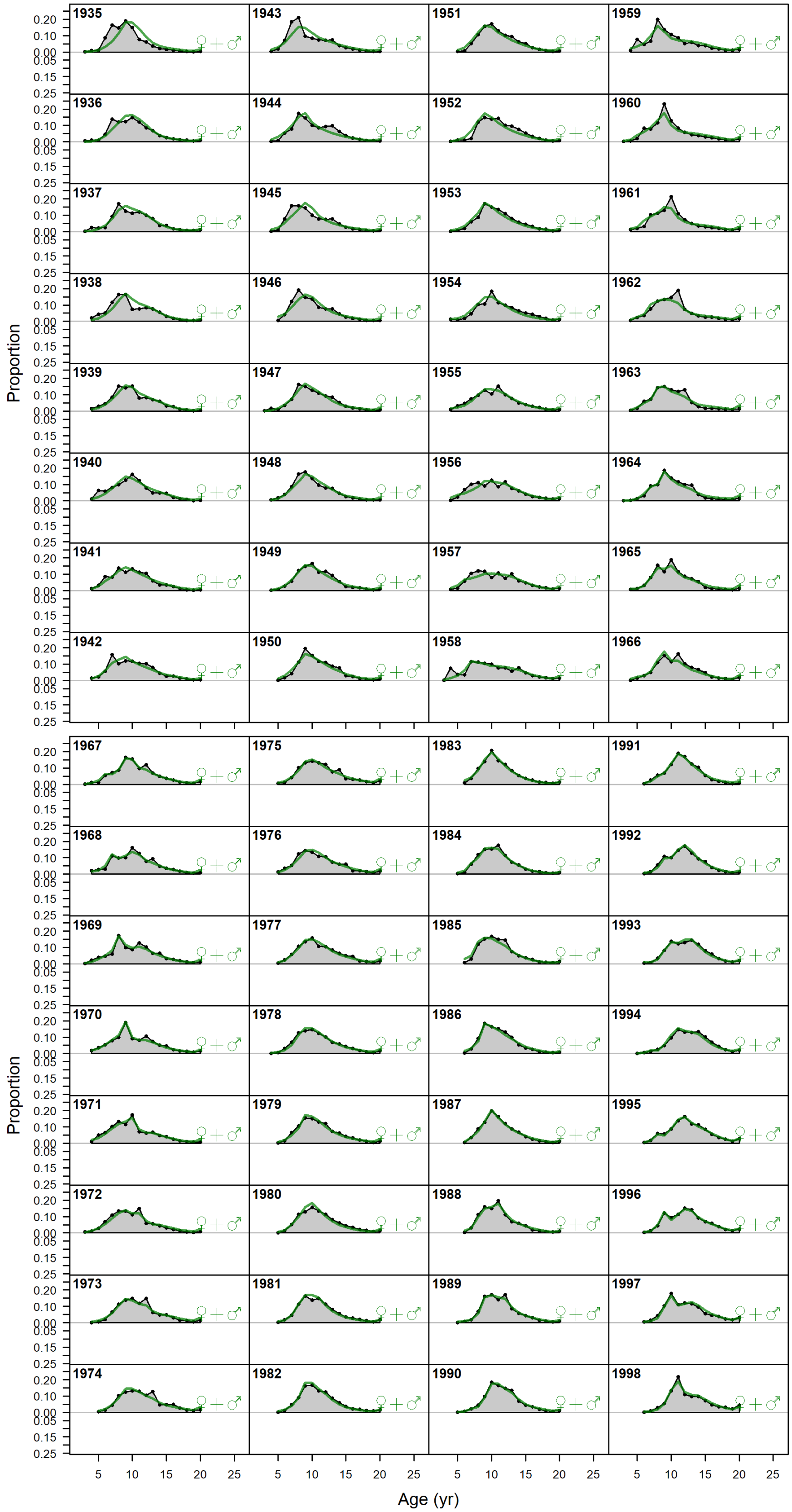


Figure 36. Fit to early years of fishery age data in the coastwide long model.

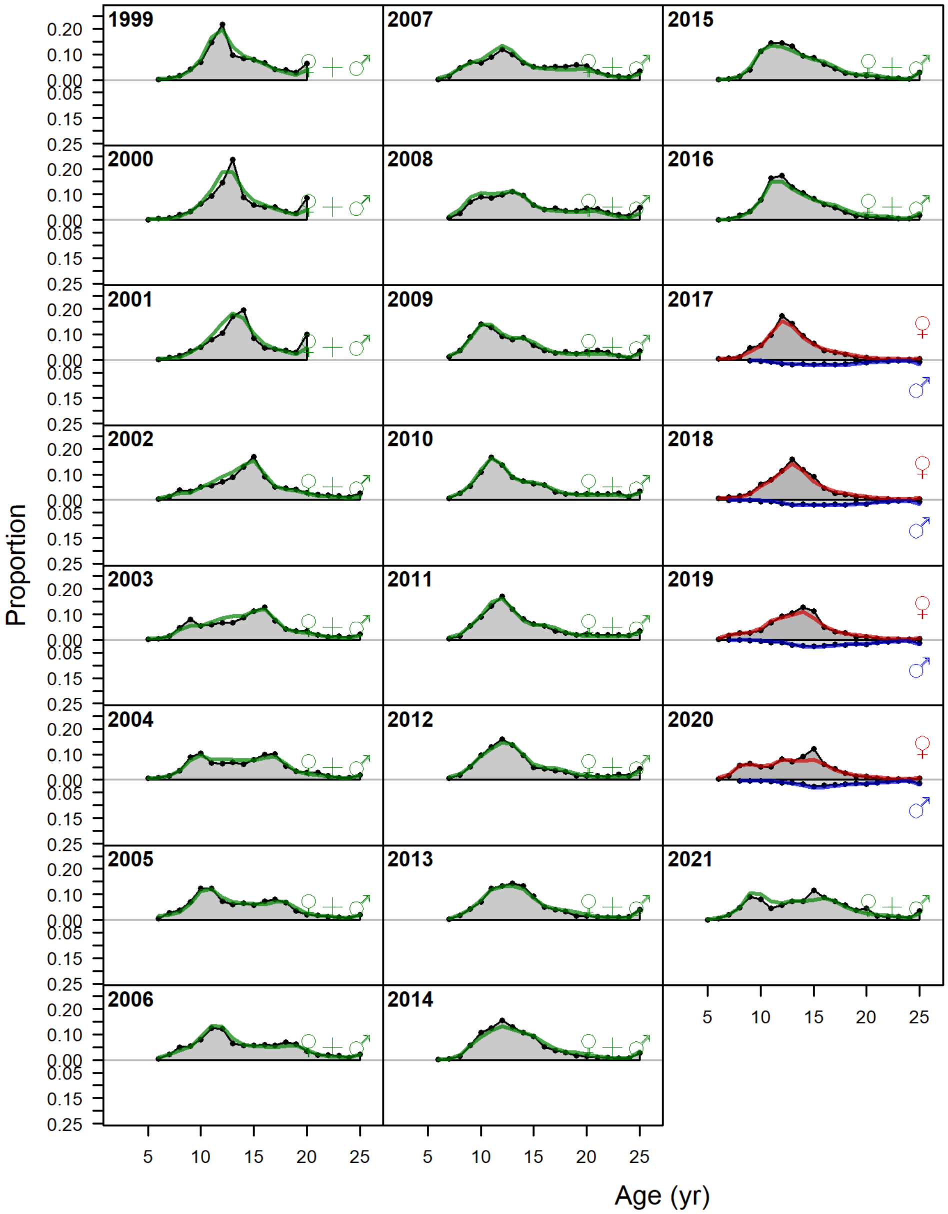


Figure 37. Fit to later years of fishery age data in the coastwide long model.

Fishery selectivity generally showed a pattern toward selecting fewer younger fish in the latter half of the time series, but a similar trend to the FISS in the most recent years (Figure 38). This may be consistent with changes in both the age-structure of the stock, the trends in size-at-age interacting with age-based selectivity and the spatial distribution creating changes in availability. Fishery catchability showed a very large (unconstrained) increase associated with the change from "J" to circle hooks (Table 17, Figure 39). Older halibut were more represented in the non-directed fishery discards age data prior to 1992, and therefore the estimated selectivity was nearly asymptotic. Recreational and discard selectivity estimates were relatively similar to those from the coastwide short model.

Female natural mortality in the coastwide long model was estimated to be higher (0.215) than for males (0.203) although the 95% intervals overlap broadly (Table 17, Figure 19). The environmental link parameter (β) was estimated to be positive (0.372), with no density below a value of 0.0, thus suggesting a strong and significant relationship between average recruitment and the phase of the PDO (Table 17). However, the time series of estimated recruitments (Figure 40) and deviates from the PDO-informed stock-recruitment relationship (Figure 41) suggested that some residual effect and/or mismatch in the relationship might still be present. Specifically, the poor PDO period from 1947-1977 and the positive phase from 1978-2006 generally correspond to negative and positive deviations even with the relationship included (Figure 41).

Summary of strengths and weaknesses for the coastwide long model:

Strengths:

- Includes uncertainty in female natural mortality
- Includes extensive historical data
- Uses environmental information to inform recruitment
- Modest technical overhead (complexity)
- Fits the fishery and survey indices well
- Fits both the survey and fishery age data well
- Internally consistent data weighting

Weaknesses:

- May lose Region-specific trends and biological patterns due to aggregation
- Relies heavily on only fishery trends over the historical period
- Implicitly assumes stationarity in some processes (e.g., the stock-recruitment function, M) over the long historical period
- Implicitly assumes that availability to the fishery did not change over the historical period, despite known patterns in geographical expansion prior to the 1960s

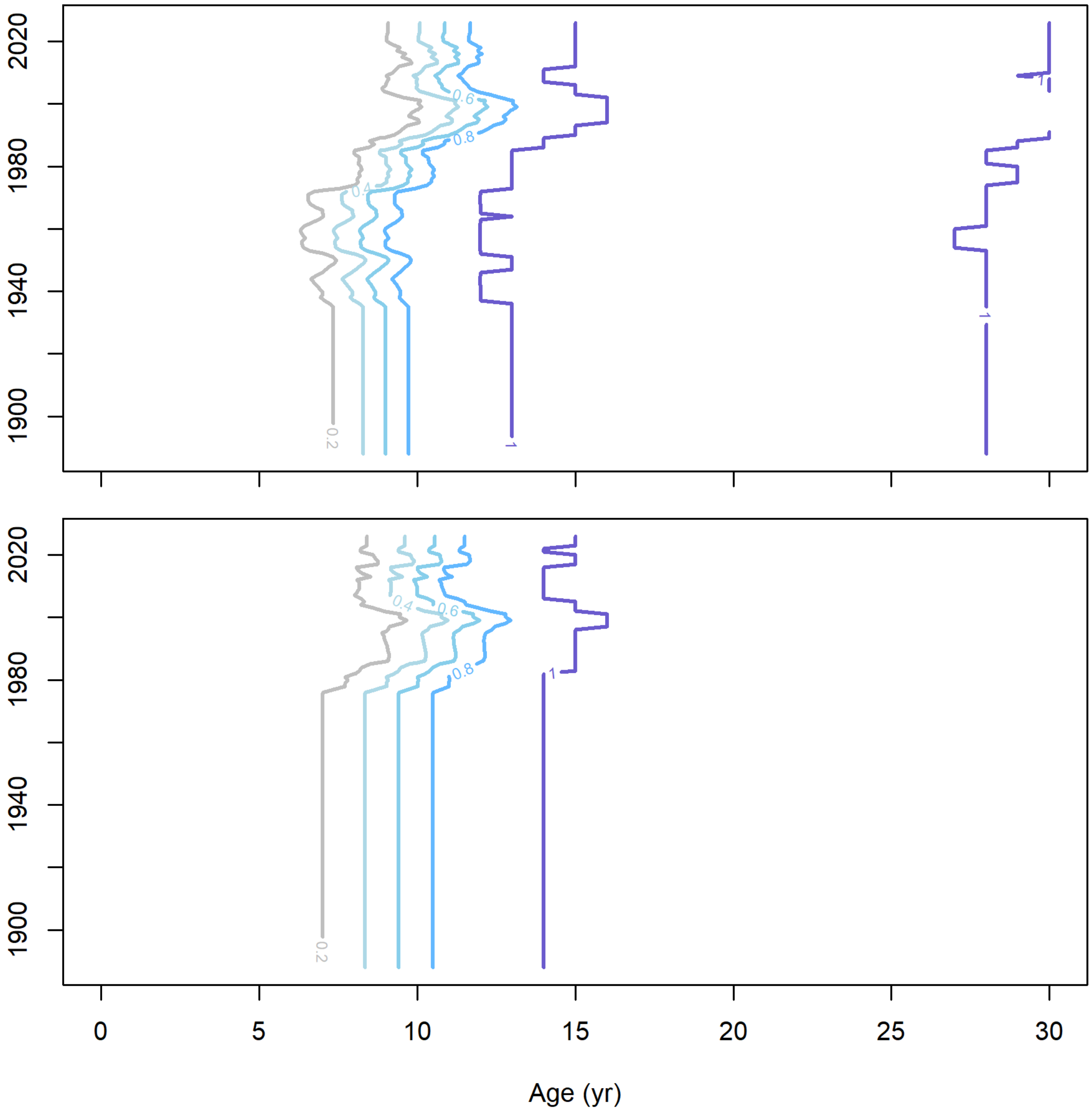


Figure 38. Estimated selectivity for females in the commercial fishery landings (upper panel) and survey (lower panel) in the coastwide long model.

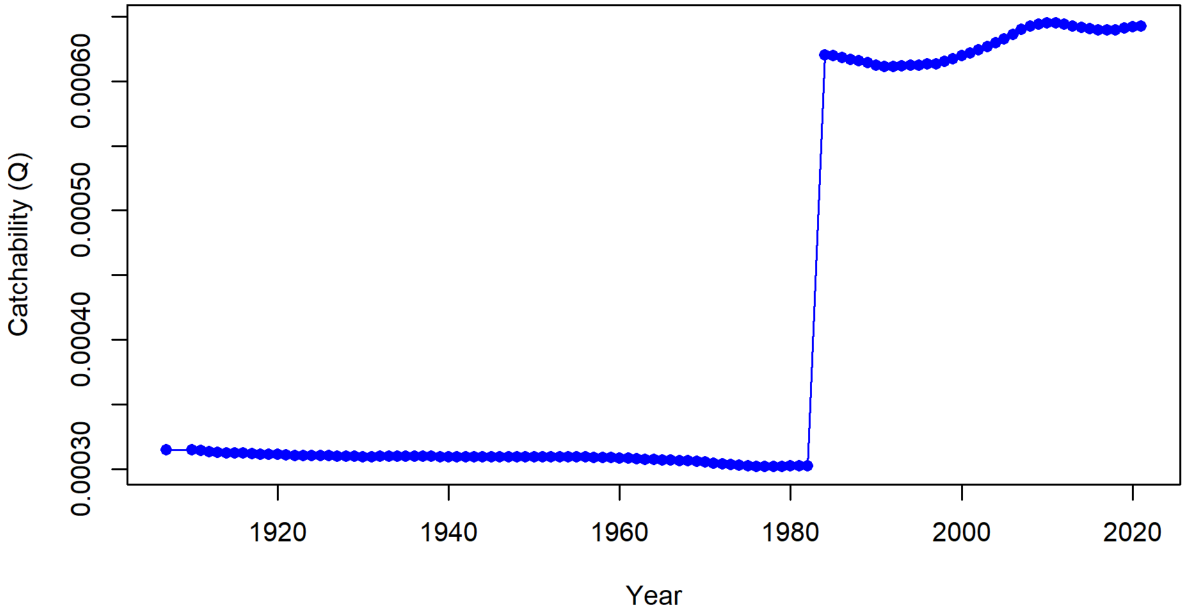


Figure 39. Time-varying fishery catchability in the coastwide long model. The change corresponding to the transition to circle hooks in 1984 is unconstrained.

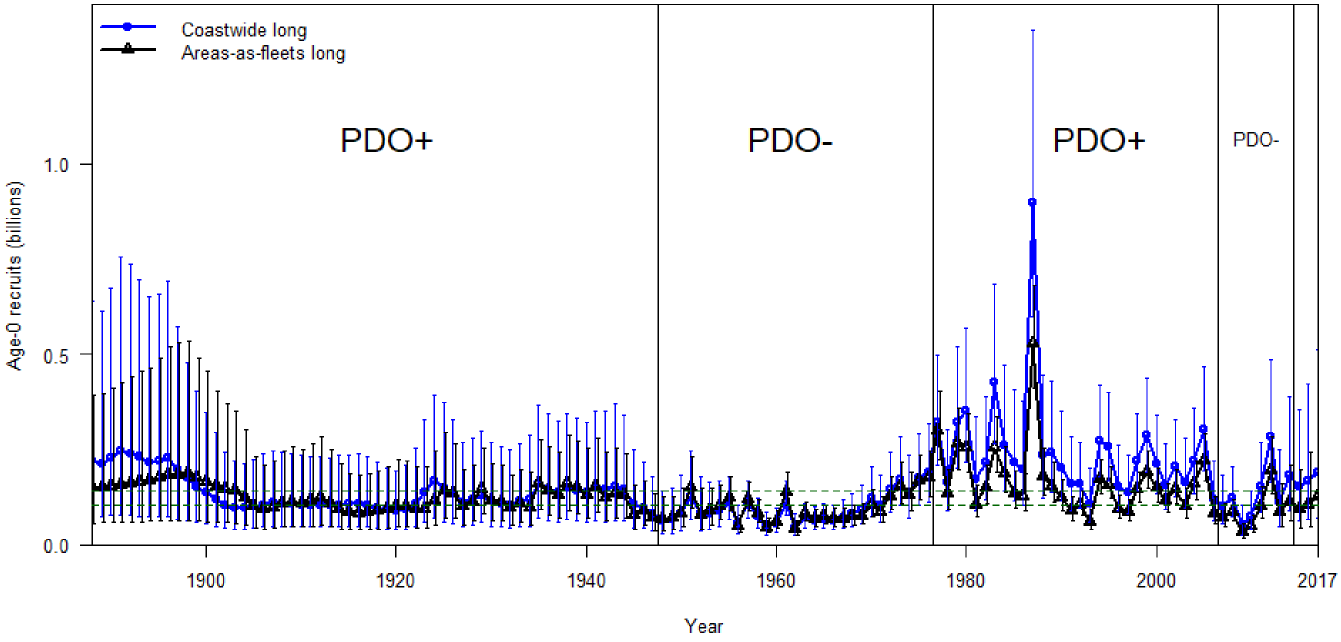


Figure 40. Estimated recruitments and assumed PDO regimes from the coastwide long and AAF long models (right panel); horizontal lines indicate equilibrium values in the absence of the PDO.

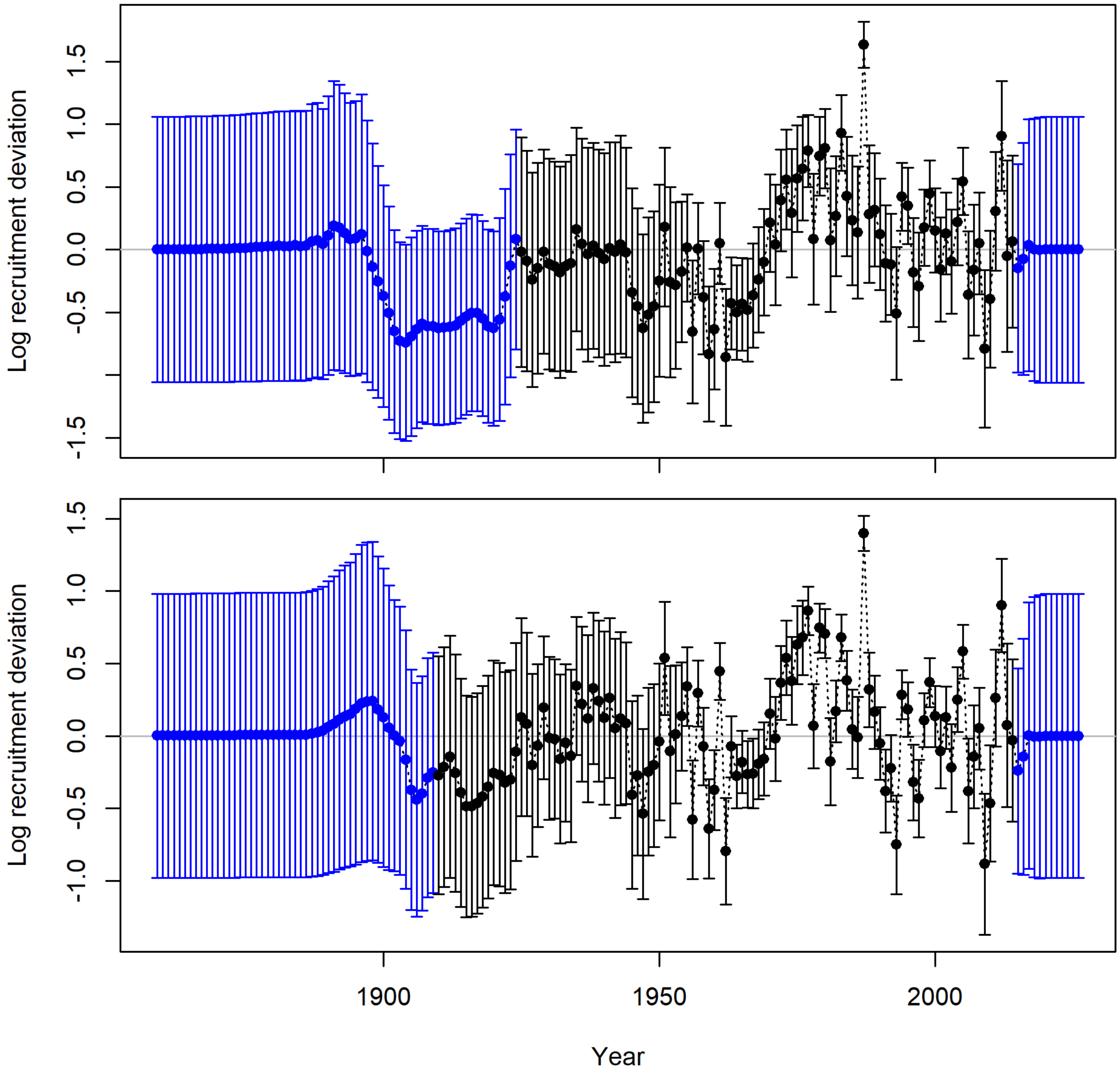


Figure 41. Estimated recruitment deviations in the coastwide long (upper panel) and AAF long (lower panel) models; horizontal lines indicate expected values based on the stock-recruitment functions as modified by the estimated PDO relationships.

AAF short

The AAF short model fit the observed trends in all fishery and FISS indices relatively well (Figure 42-42). Fit to the aggregate age data for each fleet clearly illustrated the differences in age structure (Figure 44). The biggest differences between female and male halibut observed from the FISS occurred in Region 3, and generally Regions 4 and 4B were predicted (and observed) to have the greatest fraction of older halibut, a majority of which were males. The fit to the annual FISS age data generally captured these patterns, with the worst fit occurring for the data from Region 2 (Figure 45); the Francis weight suggested a low weighting for the Region 2 FISS data consistent with these patterns (Table 16). Although showing a reasonably good aggregate fit, the fit to annual commercial fishery landings in Biological Regions 4 and 4B (Figure 46-46Figure 47) did not capture the strong peaks created by the 1987 year-class in the late 1990s and early 2000s; however of these fleets only the Region 4 data were down-weighted from the bootstrapped inputs based on the Francis weighting (Table 16). No model configurations evaluated during model development were able to fit the peak observations of this cohort observed in Regions 4 and 4B, which may be a reflection of the spatial nature of the dynamics not well approximated by an AAF approach.

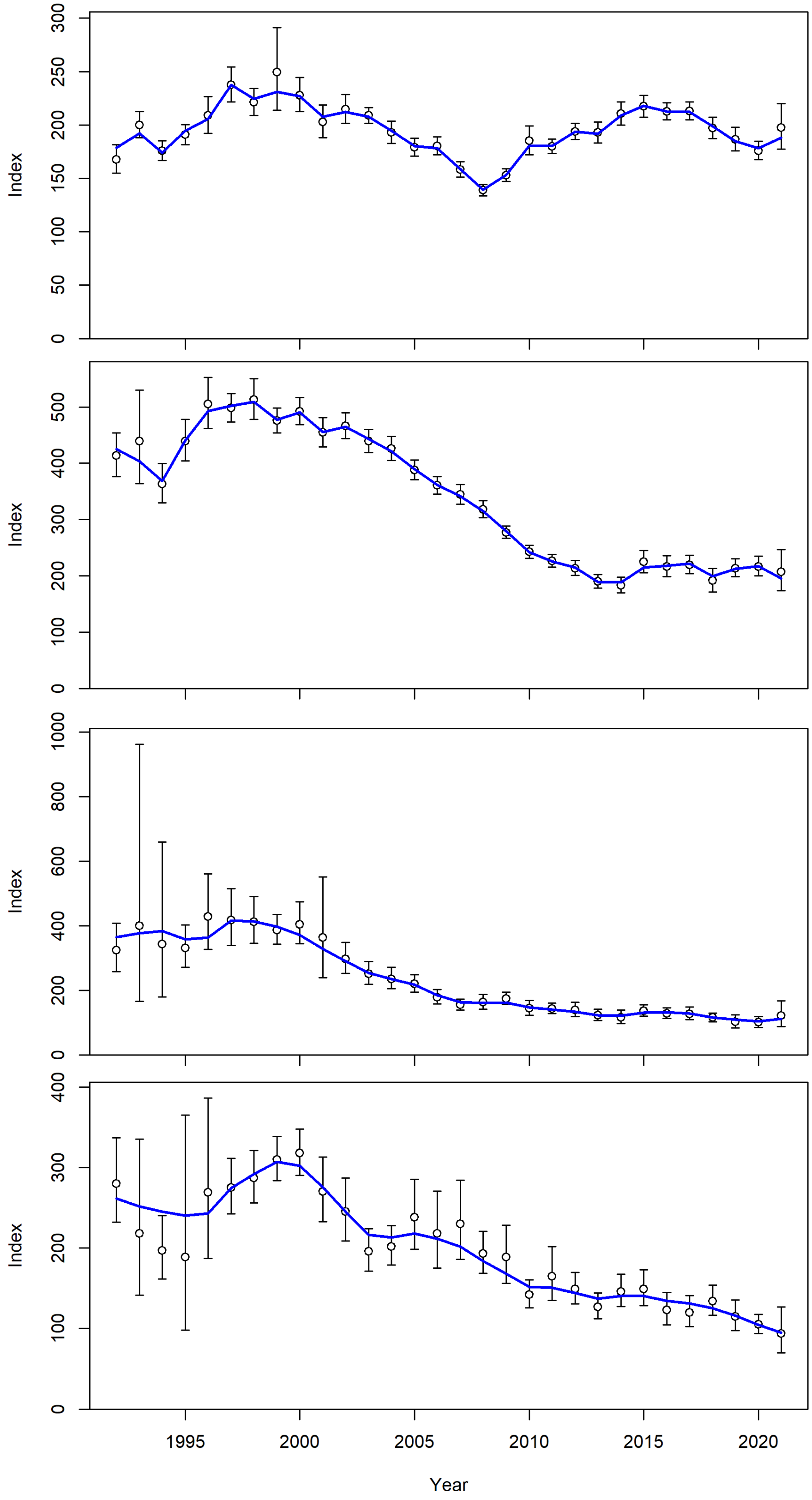


Figure 42. Fit to fishery trends in Biological Regions 2, 3, 4, and 4B (top to bottom) in the AAF short model.

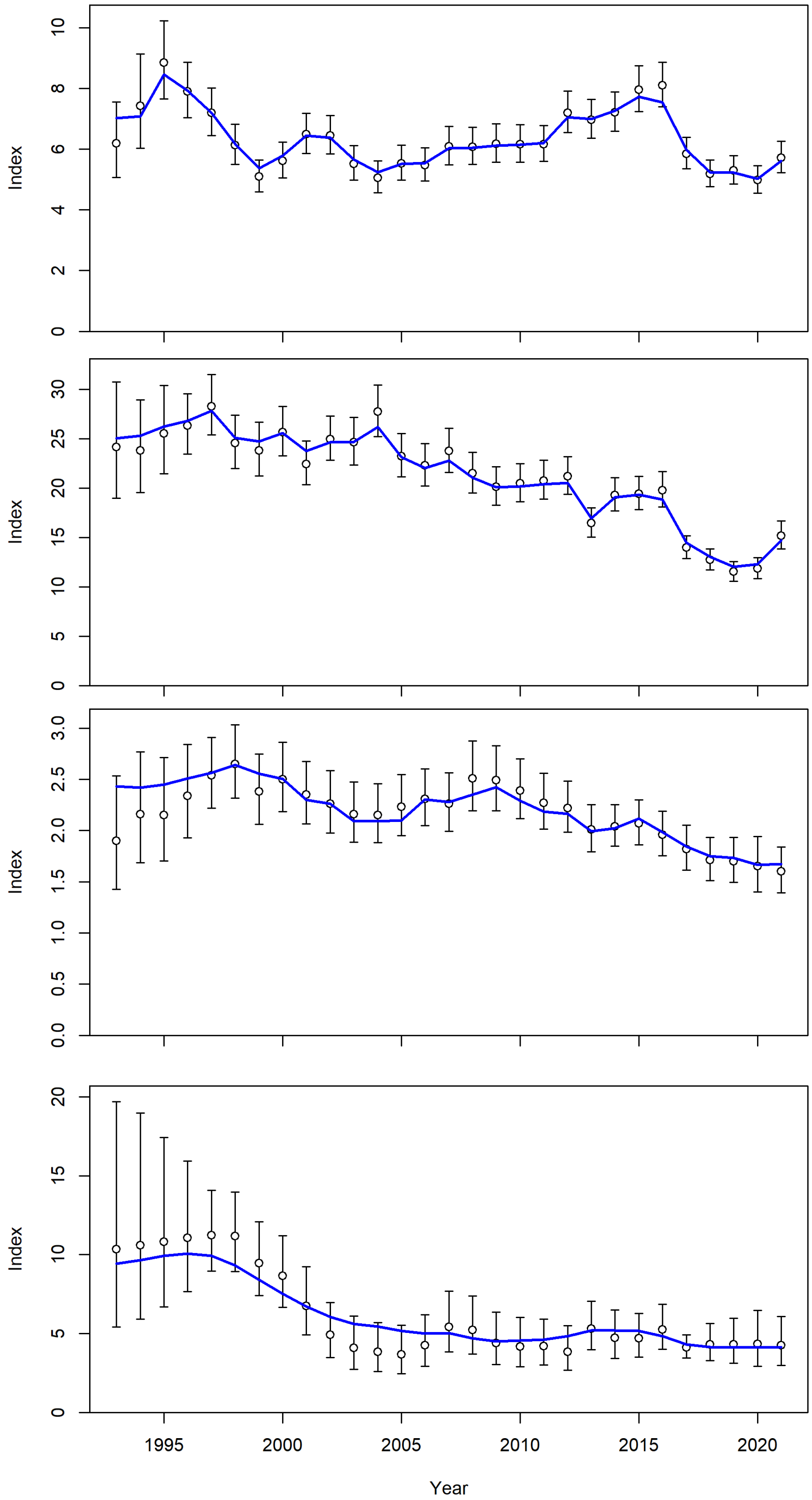


Figure 43. Fit to survey trends in Biological Regions 2, 3, 4, and 4B (top to bottom) in the AAF short model.

The estimate of female natural mortality in the AAF short model (0.211) was slightly lower than in the coastwide long model and male value much lower (0.177; Table 17). The lack of overlap on the 95% intervals indicates the clearly different explanation in this model for the observed sex-ratios, albeit restricted to the most recent portion of the time-series. This result likely indicates the trade-off between the assumption of asymptotic selectivity in the coastwide model and domed selectivity for most Regions in the AAF models. The AAF short model estimated a negative but somewhat smaller initial offset to recruitment. Due to the higher estimated *M*, the AAF short model estimated a higher absolute level of recent recruitment and spawning biomass than the coastwide short model (Table 17).

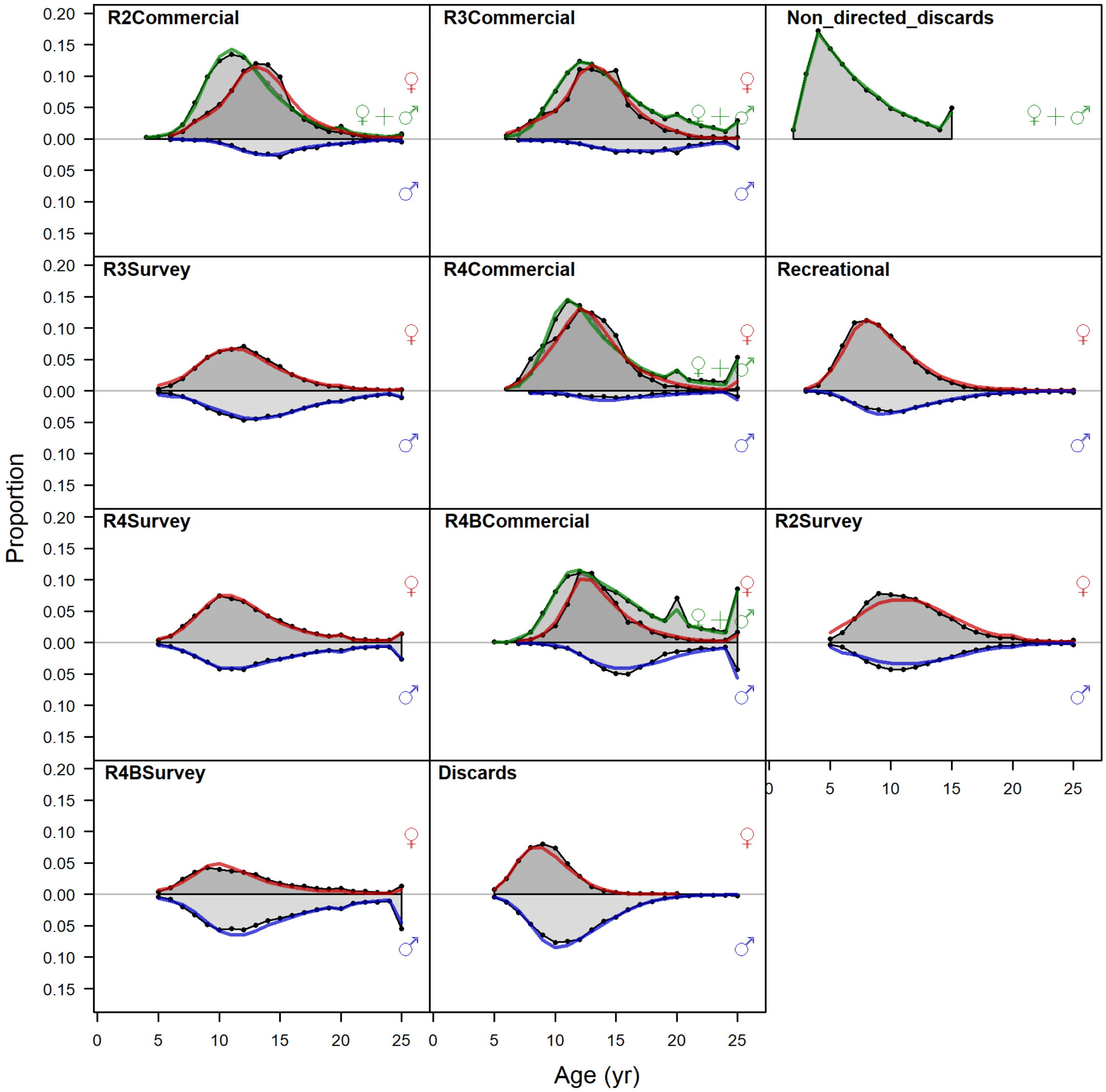


Figure 44. Aggregate fit to age data for each model fleet in the AAF short model.

Summary of strengths and weaknesses for the AAF short model:

Strengths:

- Parameter estimates are derived from the most recent time period
- Avoids aggregating data over Biological Regions with differing trends and biological patterns
- Fits the Regional fishery and FISS indices well
- Fits Regions 2 and 3 fishery age data well
- Internally consistent data weighting
- Propagates uncertainty in female and male *M* estimates

Weaknesses:

- Does not include environmental information to inform recruitment
- Modest technical overhead (complexity)
- Residual patterns in Region 4 and 4B fishery and survey age data
- Fits Regions 2 and 3 FISS age data poorly
- Does not include extensive historical data

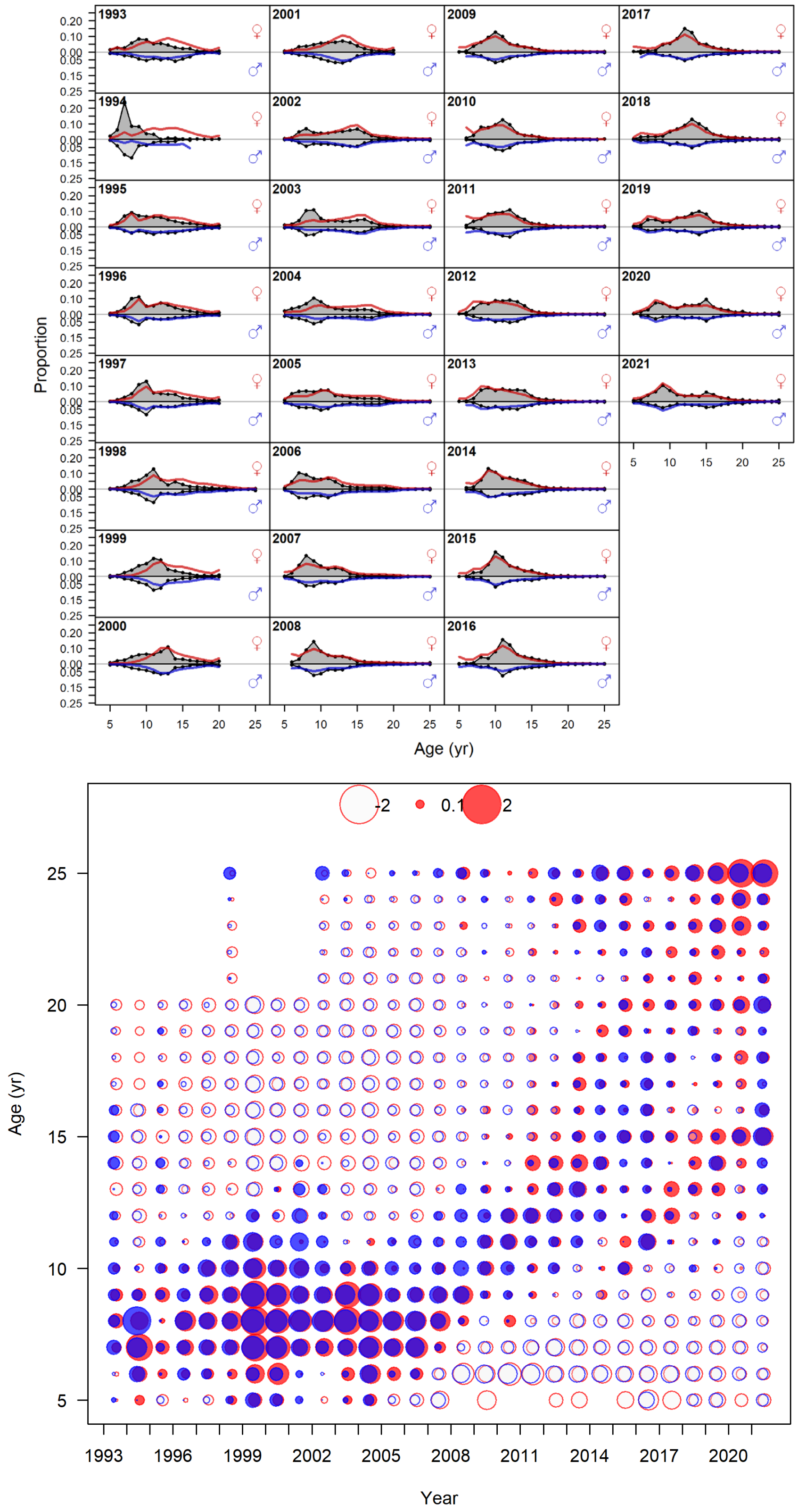


Figure 45. Fit to age data (upper panel) and Pearson residuals (lower panel) from the Region 2 FISS in the AAF short model; red circles denote female residuals, and blue circles denote male residuals.

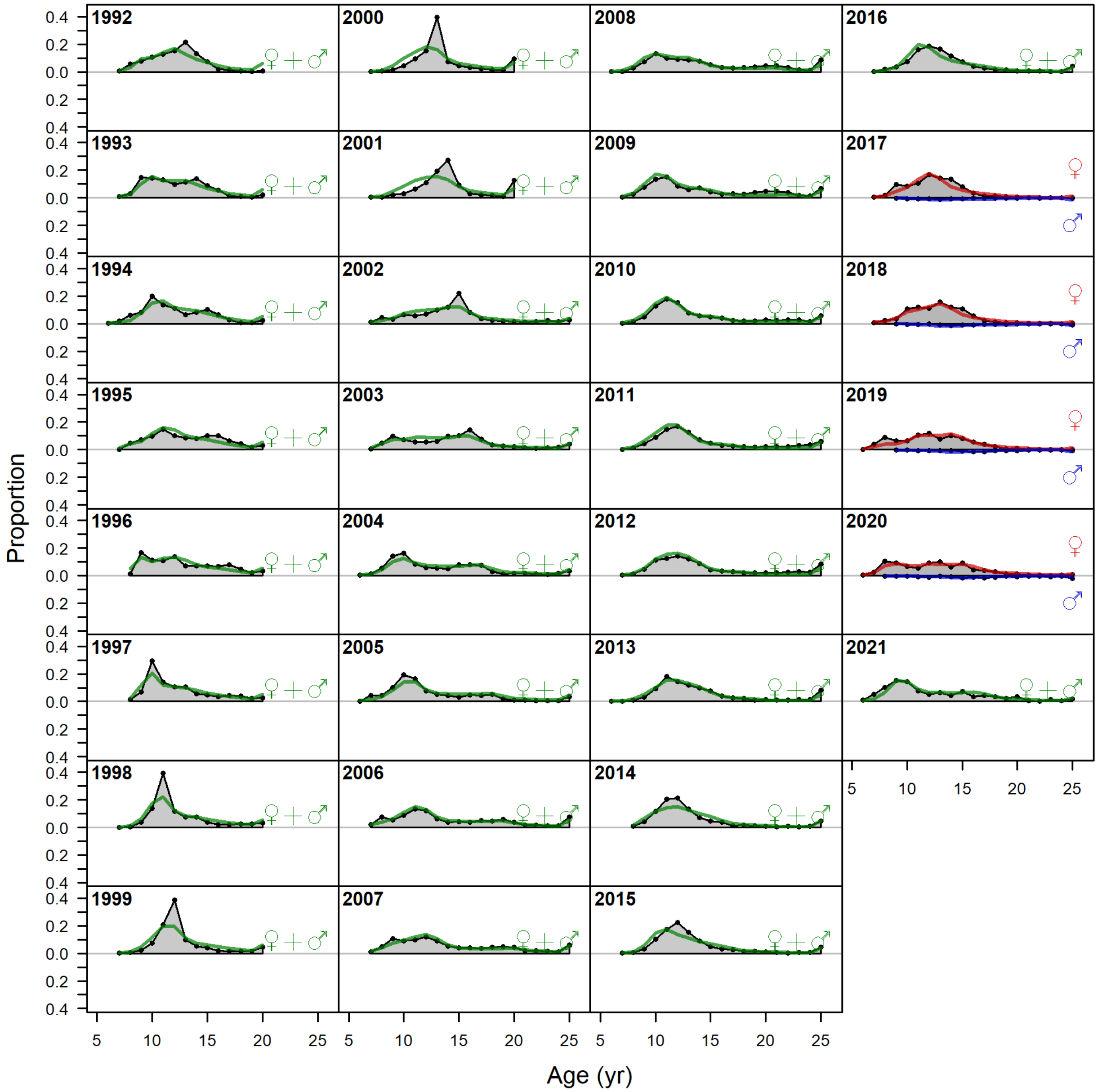


Figure 46. Fit to age data from the Region 4 commercial fishery landings in the AAF short model.

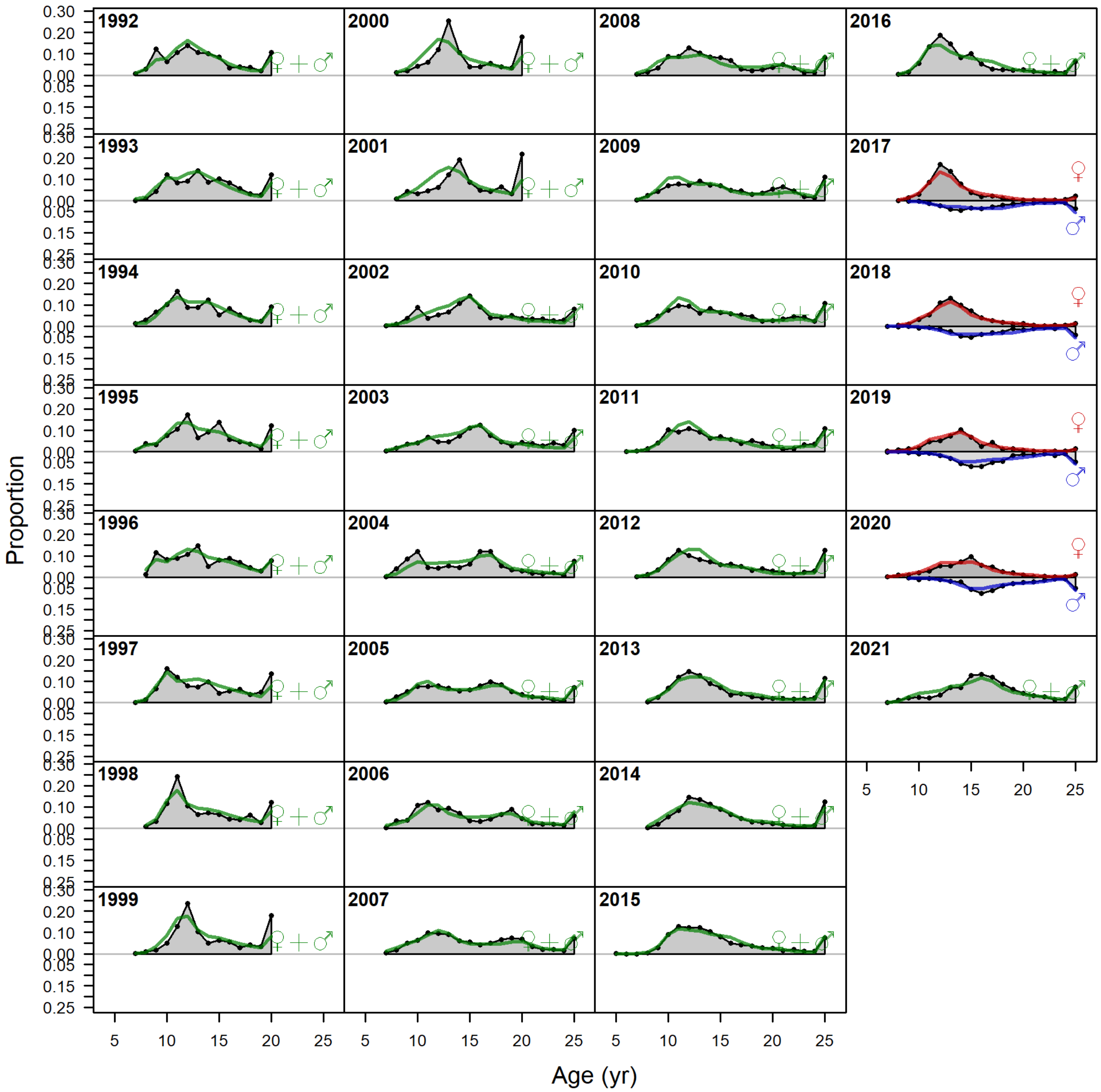


Figure 47. Fit to age data from the Region 4B commercial fishery landings in the AAF short model.

AAF long

Like the AAF short model, the AAF long model fit both the fishery and FISS trends well (Figure 48-48). Aggregate fits to the FISS age composition data showed similar patterns to those observed in the AAF short model (Figure 50). The fit to the FISS age data improved over the time series, but the Region 2 and 3 FISS age data was strongly down-weighted in order to achieve consistency with the Francis weighting (Table 16). This resulted in the worst fit by fleet (Figure 51-52). Lack of fit to the Region 3 FISS data occurred primarily in the early part of the time-series Figure 52. Among the fishery fleets, the Region 4 data were most heavily down-weighted from the bootstrapped input sample sizes (Table 16).

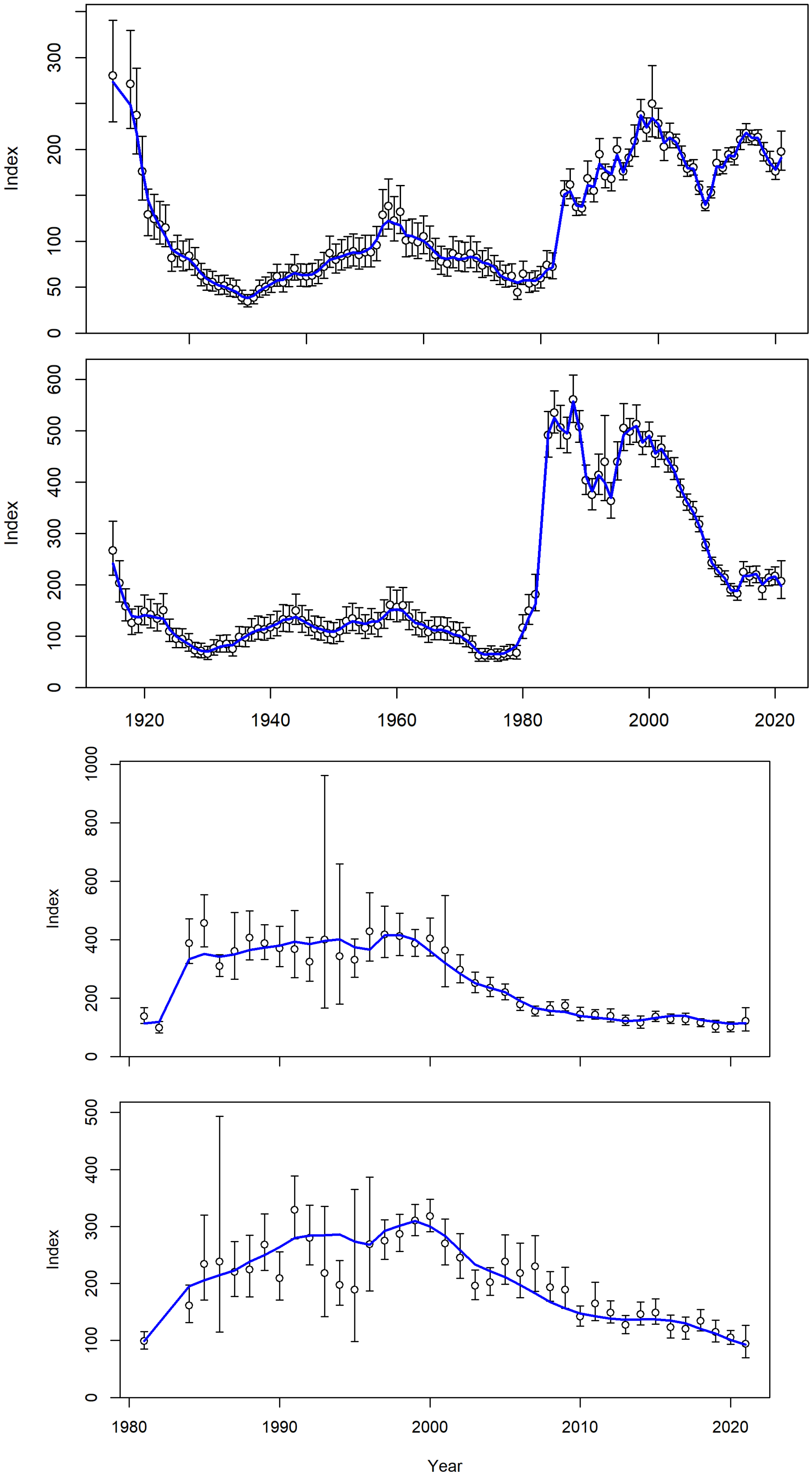


Figure 48. Fit to fishery trends in Biological Regions 2, 3, 4, and 4B (top to bottom) in the AAF long model.

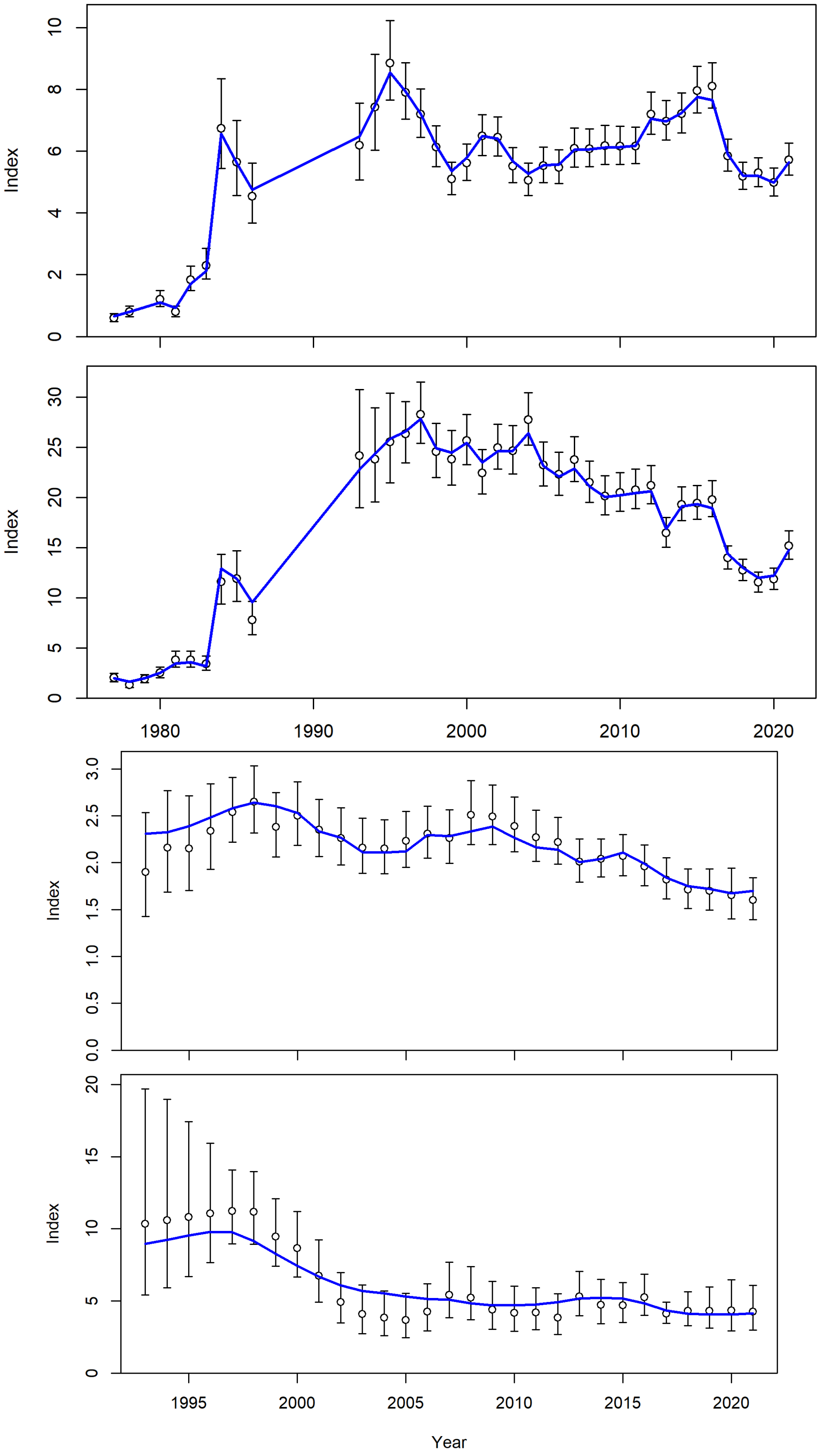


Figure 49. Fit to FISS trends in Biological Regions 2, 3, 4, and 4B (top to bottom) in the AAF long model.

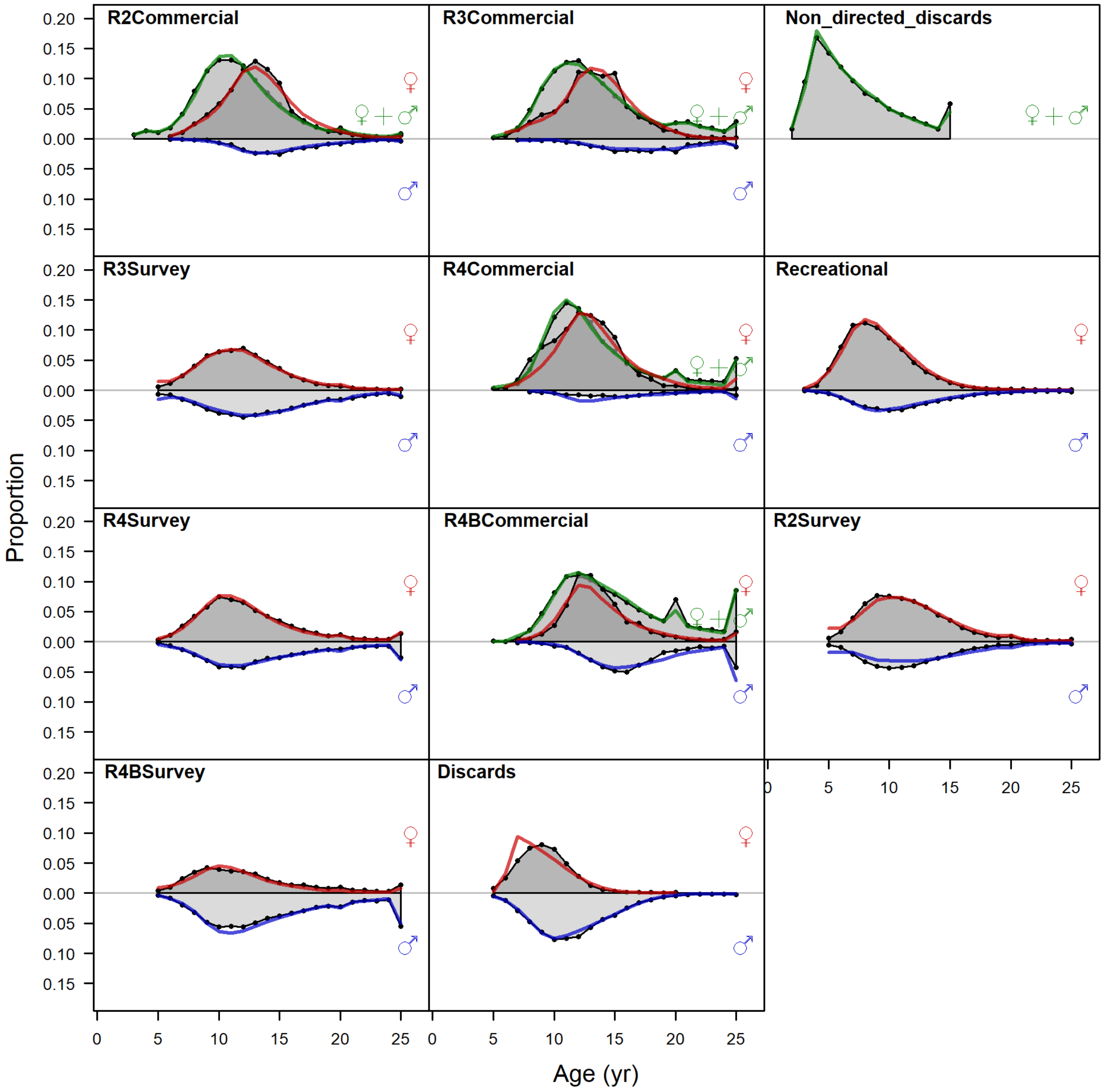


Figure 50. Aggregate fit to age data for each model fleet in the AAF long model.

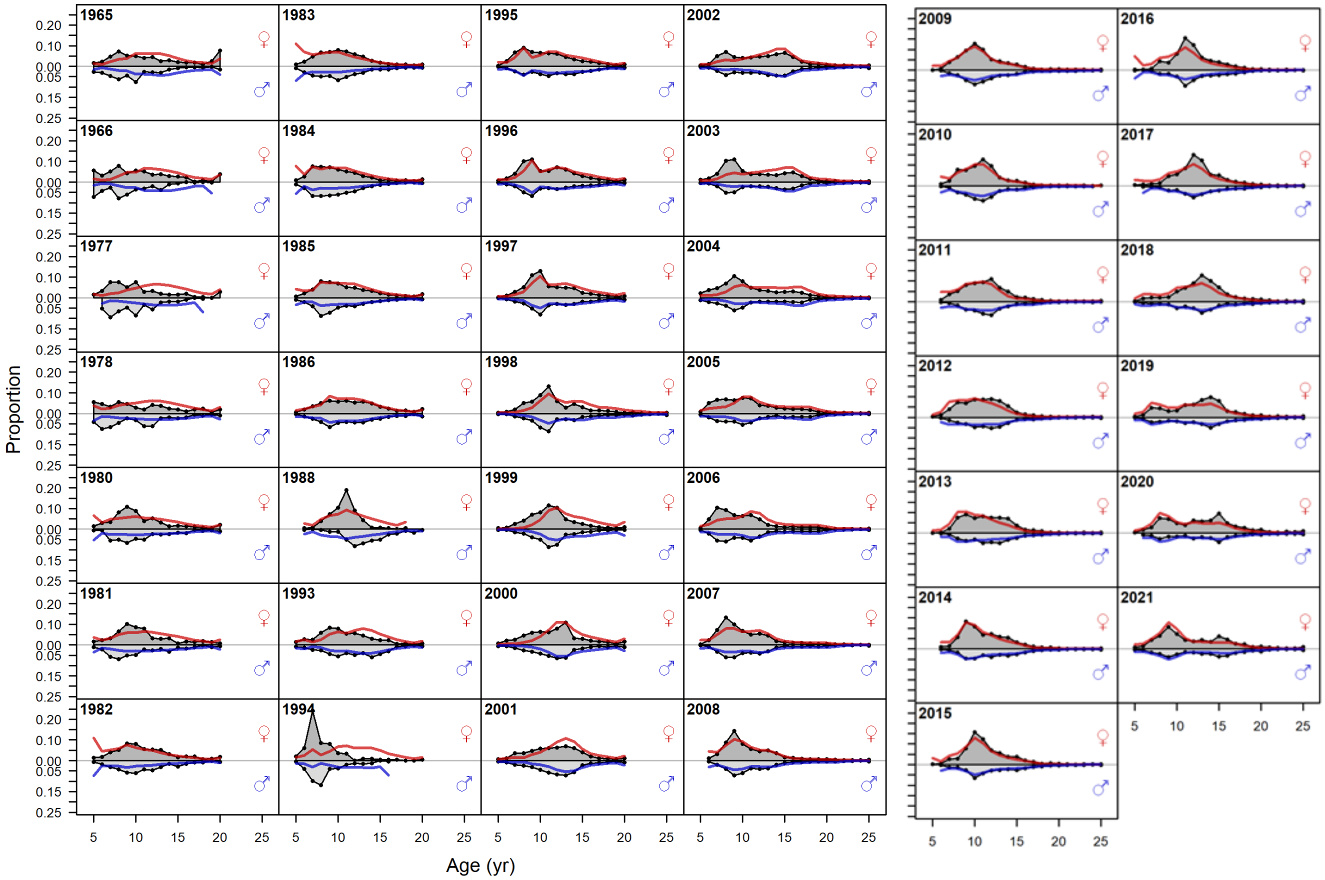


Figure 51. Fit to age data from the Region 2 FISS in the AAF long model.

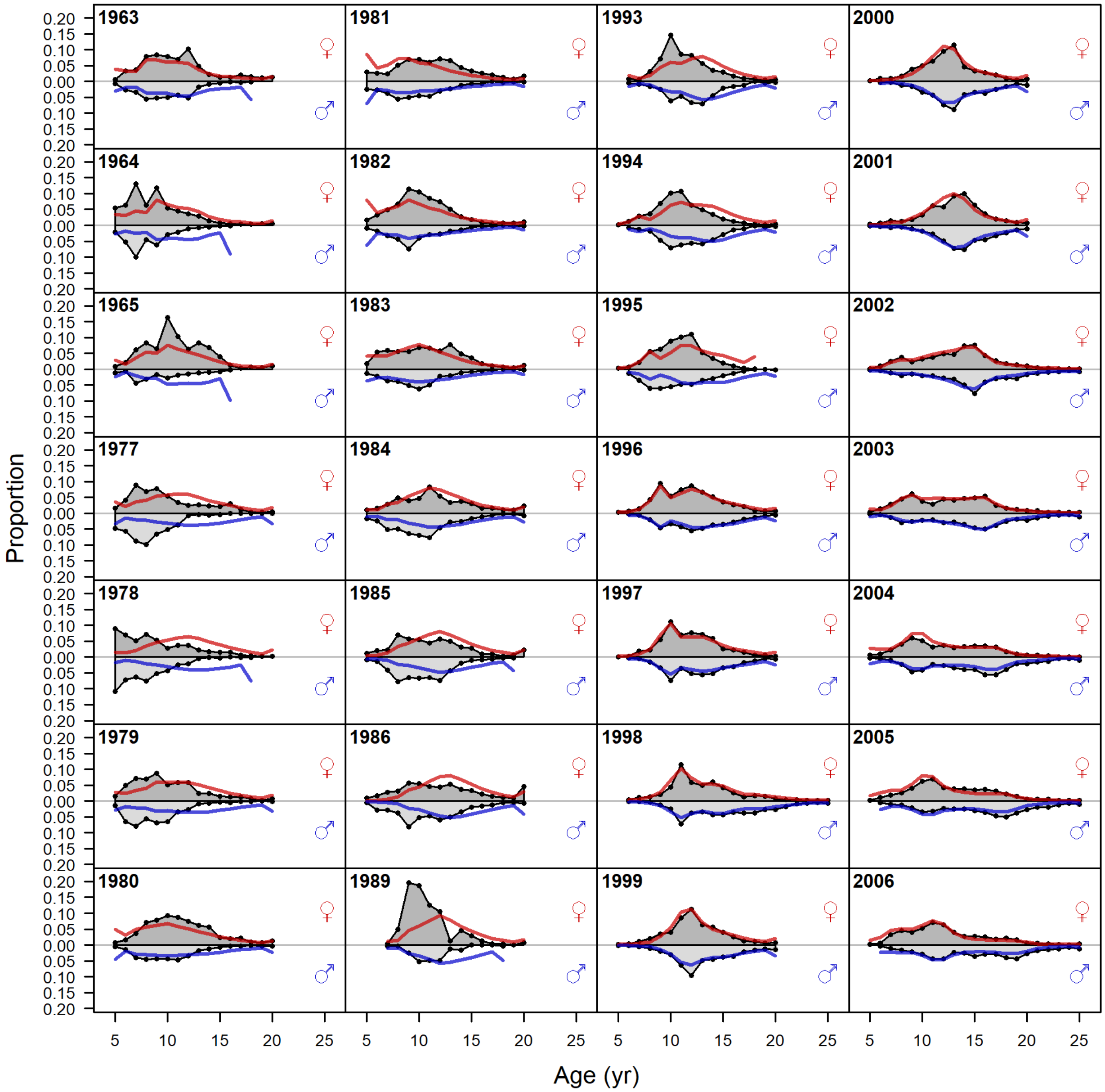


Figure 52. Fit to early age data from the Biological Region 3 FISS in the AAF long model.

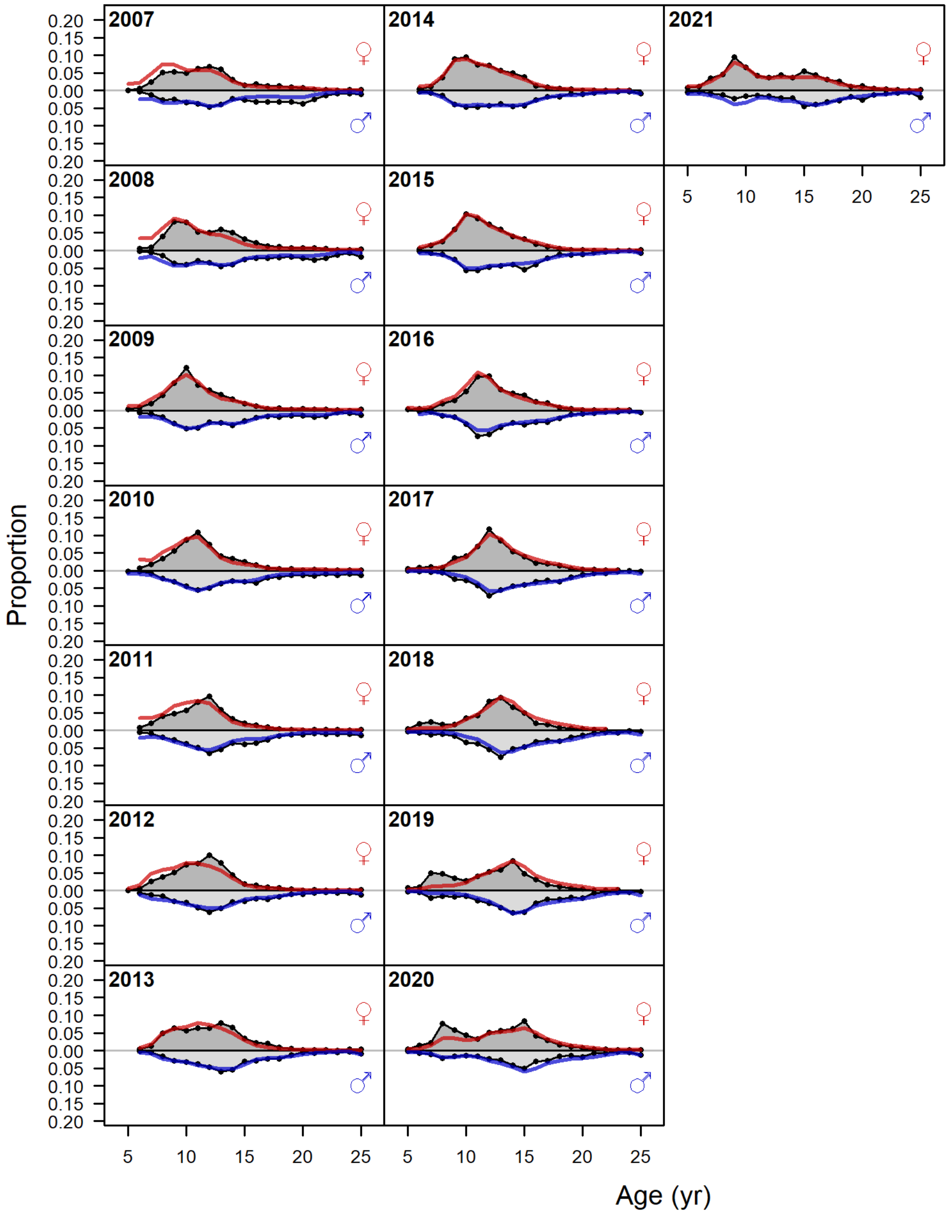


Figure 53. Fit to later age data from the Biological Region 3 FISS in the AAF long model.

Similar to the AAF short model, FISS selectivity was estimated to be asymptotic for Biological Regions 4 and 4B. Peak male selectivity in the commercial fishery landings was also estimated to be asymptotic. All fleets with data extending past the transition from J to circle hooks in 1984 showed a strong offset in the unconstrained deviation in catchability for that year (Table 17). Discard and recreational selectivity estimates were similar in the AAF long model to those estimated in the coastwide long model. Non-directed discard selectivity was estimated to be domed, again illustrating the trade-off between domed fleets in the AAF models and asymptotic selectivity over the entire time-series in the coastwide models. This likely interacts with the estimation of natural mortality, producing slightly lower values in the AAF long model (0.184 for females, and 0.164 for males) than in the coastwide long model (Table 17).

The environmental link coefficient was estimated to be slightly weaker (0.349) than in the coastwide long model, although the 95% interval did not contain zero (Table 17).

The AAF long model produced intermediate estimates of recent recruitment and female spawning biomass (Table 17). This result is consistent with the intermediate estimates of male and female *M* from this model.

Summary of strengths and weaknesses for the AAF long model:

Strengths:

- Includes uncertainty in female and male *M*
- Includes extensive historical data
- Uses environmental information to inform recruitment
- Fits the fishery and survey indices well
- Fits both the Regions 2, 3 and 4B fishery age data well
- Fits Region 4 and 4B FISS age data well
- Internally consistent data weighting

Weaknesses:

- Highest technical overhead (complexity) of the four models
- Most challenging model to check and ensure reliable convergence
- Relies heavily on only fishery trends over the historical period
- Implicitly assumes stationarity in some processes (e.g., the stock-recruitment function, *M*) over the long historical period
- Fit Biological Regions 2 and 3 survey age data poorly

Sources of uncertainty

The four models evaluated here represent, within the set itself, significant sources of uncertainty in how to treat the data (partitioning by fleets or aggregating to a single series), as well as how to treat the time-series (emphasizing the recent dynamics or including more historical information). Further, the differing assumptions of fixed vs. estimated female natural mortality rate is also embedded in the differences observed among the four model results. These factors lead to differences in both scale and trend. In aggregate, the four models together reflected much more uncertainty than any single model, while still showing a similar basic trend over the recent time-series’ of both spawning biomass and recruitment.

Sensitivity analyses

Many alternative model configurations were evaluated during model development, but only a subset of these is reported here. Several of the bridging steps from the 2021 models to the 2022 preliminary models also represent sensitivity analyses. Further, the section below providing likelihood profiles over female *M* clearly illustrates *M* as one of the largest uncertainties in this assessment. Sensitivity analyses specifically intended to highlight the importance of ongoing research (e.g., whale depredation, maturity curves, etc.) are produced each year as part of the final stock assessment (Stewart and Hicks 2022).

The large differences in the scale of the spawning biomass in the historical period between the two long time series models represent the range of assumptions about the connectivity of the stock via spatial availability (Figure 54). Specifically, domed selectivity for Biological Regions 2 and 3 in the long AAF model implicitly assumes that older fish (located in northern and western areas) were historically less available and therefore not mobile enough to be readily available to those fisheries. Conversely, in the coastwide long model the assumption of asymptotic selectivity implies a high degree of availability and therefore connectivity between all geographic components in the population. Sensitivity analyses in the 2015 assessment indicted that these two models could be made much more similar by adjusting the degree of domed selectivity (Stewart and Martell 2016). The use of both models encompasses the range of uncertainty that exists over this aspect of the historical population dynamics, thus the primary sensitivity in the stock assessment is included in the ensemble results. A similar and consistent approach is employed to capture this dimension of uncertainty in the MSE operating models.

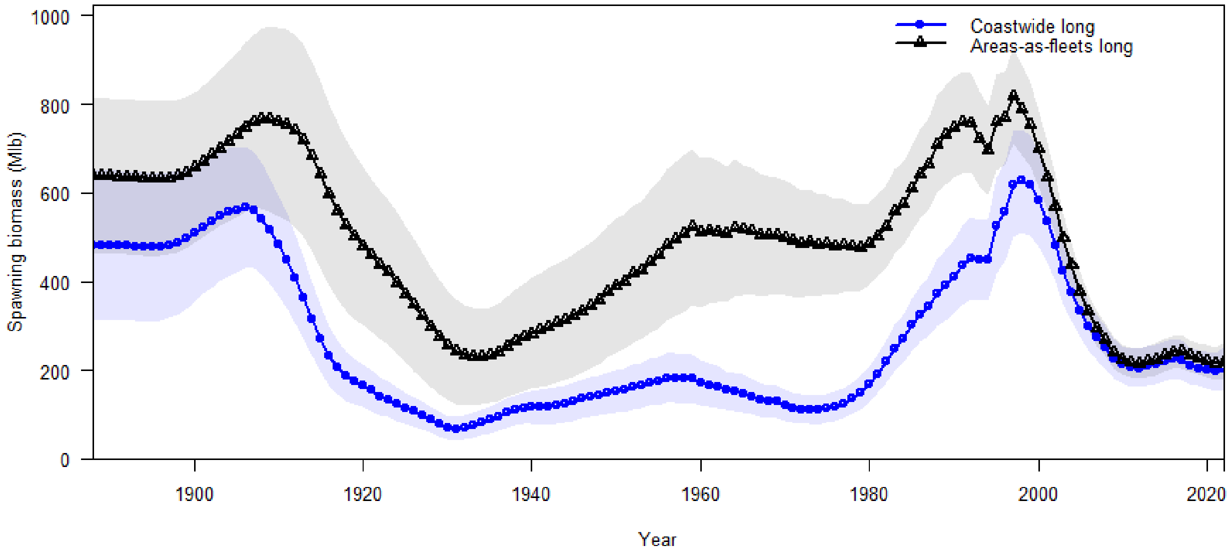


Figure 54. Comparison of the spawning biomass for the long coastwide and AAF models.

The specific technical treatment of the PDO in the two long time-series models has been identified as a research priority (IPHC 2021) and was explored extensively for this preliminary assessment. As described above, the current approach classifies the PDO into a series of binary 'regimes', and then estimates a coefficient describing the effect of these regimes on the equilibrium recruitment used in the stock-recruitment relationship. To the degree that there is still considerable variability remaining in the annual recruitment deviations, it is possible that alternative, or additional covariates might provide a similar or better explanation for observed recruitment variability.

It is common to test a wide range of possible covariates at different spatial and temporal scales. However, this approach may easily lead to false-positive relationships as the number of covariates can easily be very large. Instead, we explored a small subset of hypotheses regarding how the PDO might be related to Pacific halibut recruitment, recognizing that other variables have been explored in the past (Clark and Hare 2002b). The five hypotheses explored were:

1. *Status quo*: Regimes as implemented in the current models capture broad trends in productivity that are correlated with Pacific halibut recruitment on average, but because of the complexity of the links and likelihood of unobserved covariates more complex treatments are not appropriate.
2. *Annual deviations*: Although potentially only a proxy for the actual factors affecting recruitment, the PDO may explain additional variability in recruitment if the annual average value (itself already a deviation, and so corrected for trends) is used directly.
3. *Effects greater than one-year but less than the full regime*: the potential for cumulative and slightly lagged effects on recruitment could suggest that a running average of the PDO might explain more of the variability than shorter or longer time-periods. A five year moving average was used.
4. *Extreme values are more sensitive than others*: If the PDO is related most to the largest recruitments and all others are generally swamped by the 'noise' in natural variability, it is possible that treating the top *X*% of observed annual average PDO values as the covariate might allow for a stronger effect size. After some initial exploration, the top 33% was used for this test.
5. *The PDO-recruitment relationship has 'broken down' or does not add explanatory power to the current models*: Excluding the PDO from a series of model runs provides a comparison for all other hypotheses.

Each of these hypotheses was implemented in both the long coastwide model and the long AAF model. The hypotheses were evaluated based on whether the Root-Mean-Squared-Error (RMSE) of the estimated recruitment deviations from the PDO-informed stock-recruitment relationship changed. An increase in the RMSE indicates a degradation in the predictive power of the stock-recruitment relationship.

Results of this sensitivity analysis indicted the status quo approach provided the best explanatory power for estimated recruitment deviations across both of the long time-series models. The RMSE of 0.42 and 0.38 for the *status quo* approach in the coastwide and AAF long models was lower than any of the other hypotheses (Table 18). The only hypothesis not directly tested was using only the top 33% of PDO observations to indicate the 'high' regime in the long AAF model – with all the additional complexity in this model, it was not able to converge reliably with only a subset of regime years informing the estimated coefficient. The various hypotheses had a relatively limited effect on the estimated time-series' of spawning biomass (Figure 55).

Table 18. Comparison of the root-mean-squared-error (RMSE) of the estimated recruitment deviations and the estimate link coefficients for the coastwide long and AAF long models under different PDO hypotheses.

	Model			
	CW long		AAF long	
Treatment of the PDO	RMSE	Coefficient	RMSE	Coefficient
Status quo (binary regimes)	0.42	0.37	0.38	0.35
Annual deviations	0.44	0.45	0.38	0.38
5-year moving average	0.45	0.34	0.39	0.32
Binary on largest 1/3rd of values	0.45	0.50	*Did not converge*	
Exclude PDO	0.48	NA	0.42	NA

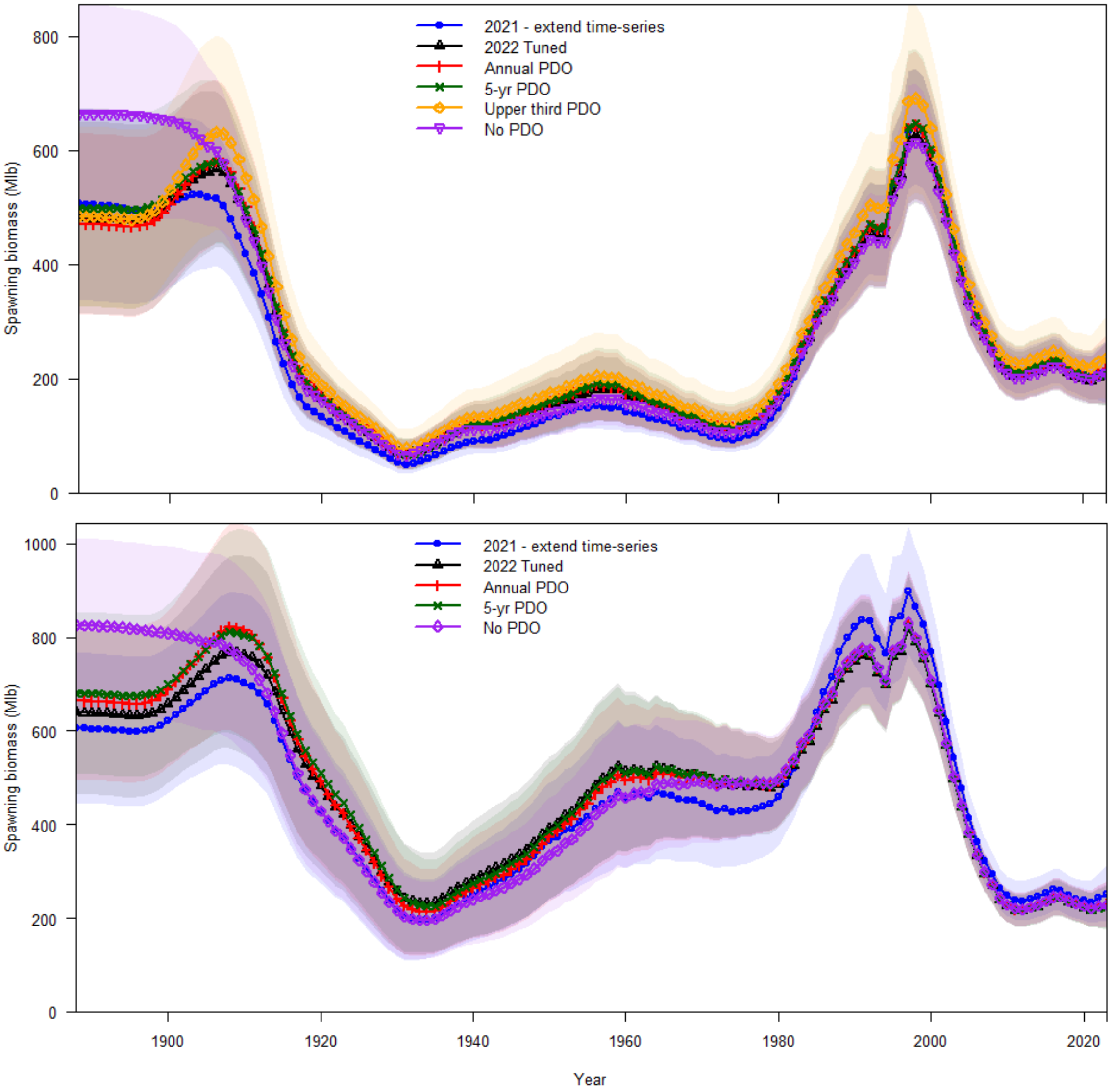


Figure 55. Comparison of the spawning biomass for the long coastwide (top panel) and long AAF (bottom panel) models across all PDO hypotheses explored.

Likelihood profiles over M

To better understand the information content of the data and the basis for estimating *M* in the Pacific halibut stock assessment a likelihood profile analysis was conducted. For each model, the value for female *M* was fixed at a series of values and all other model parameters were re-estimated. Negative Log-Likelihood (NLL) values (including the informative prior on *M*) were recorded for each fixed value of female *M* ranging from well below to well above the range included in the current models (0.1 to 0.25).

Results of the likelihood profiles indicated that the data in all four model configurations showed strong support for the upper end of the range considered. Specifically, the short coastwide model, as in all recent assessments, did not identify a minimum over the range explored (nor for several vales higher than those reported here (Figure 56). The age data, recruitment penalty and initial recruitment penalties all contributing to the higher NLL at lower female *M* values. In contrast, all three of the other models showed a minimum in the NLL informed by the same data sources and model penalties (Figure 57-58). The likelihood surface for the AAF long model was clearly irregular, illustrating that there were multiple similar parameter combinations for *M* values, particularly those below the current MLE (Figure 59). The coastwide long and AAF short models had a similar likelihood profile, with no indication that *M* was more poorly estimated in the AAF short model than in the coastwide long model where it has been reliably estimated for years.

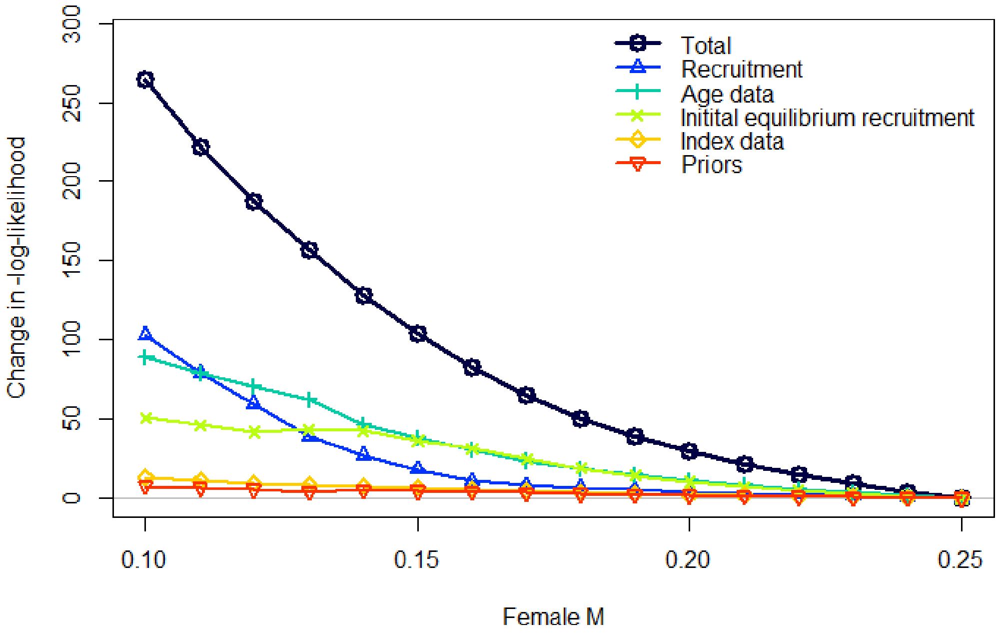


Figure 56. Likelihood components from the likelihood profile on female *M* from 0.10 to 0.25 for the coastwide short model.

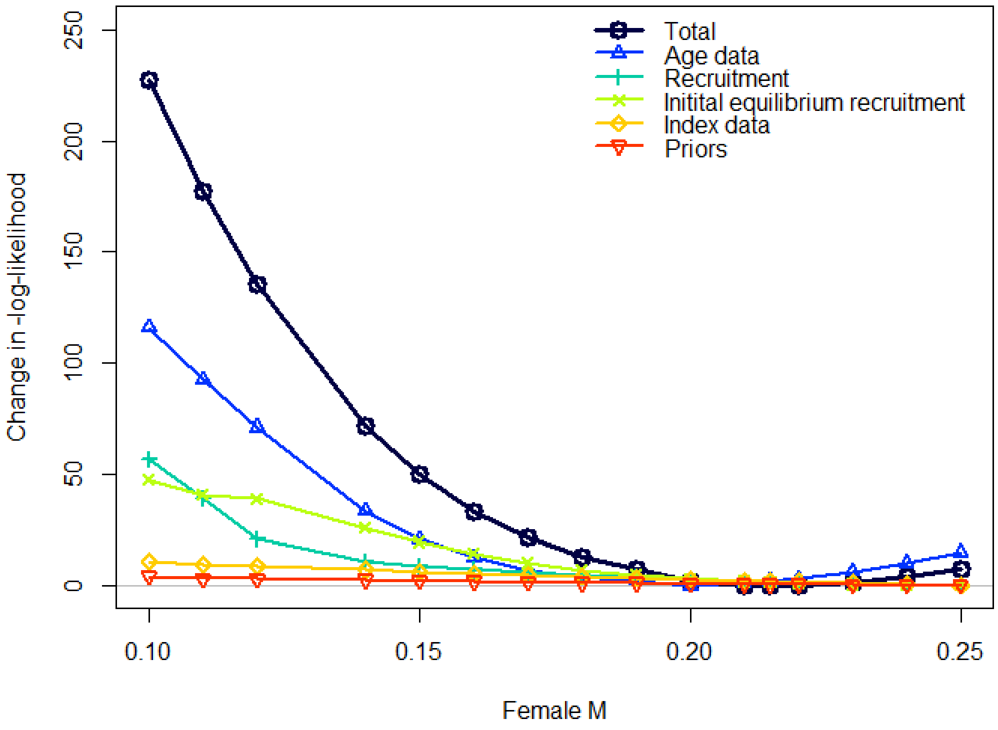


Figure 57. Likelihood components from the likelihood profile on female *M* from 0.10 to 0.25 for the AAF short model.

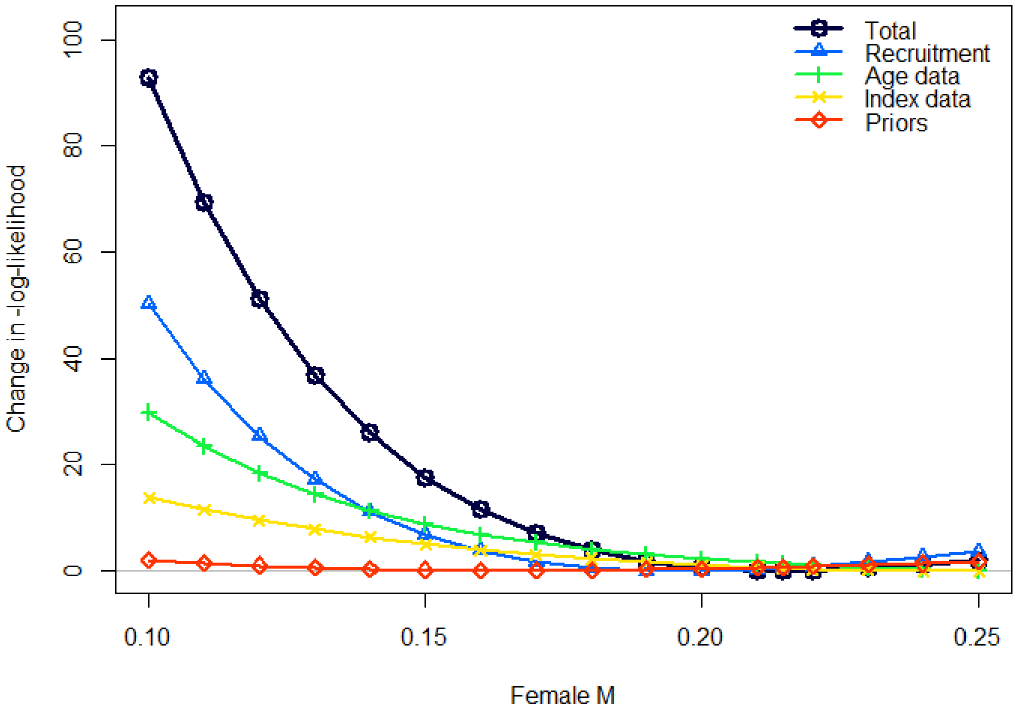


Figure 58. Likelihood components from the likelihood profile on female *M* from 0.10 to 0.25 for the coastwide long model.

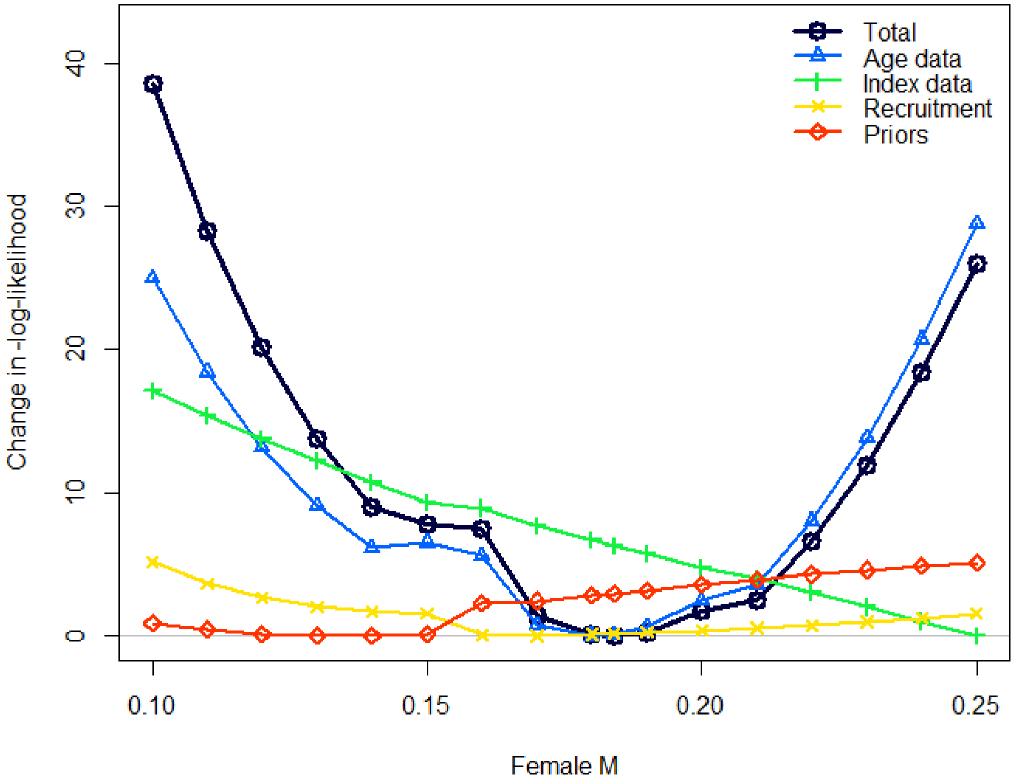


Figure 59. Likelihood components from the likelihood profile on female *M* from 0.10 to 0.25 for the AAF long model.

Similar to most fisheries stock assessments the value of *M* used in the model is closely correlated with stock productivity, and for Pacific halibut absolute size of the estimated spawning biomass. For all four models, larger values of female *M* corresponded to larger values of spawning biomass across the entire time-series (Figure 60-62).

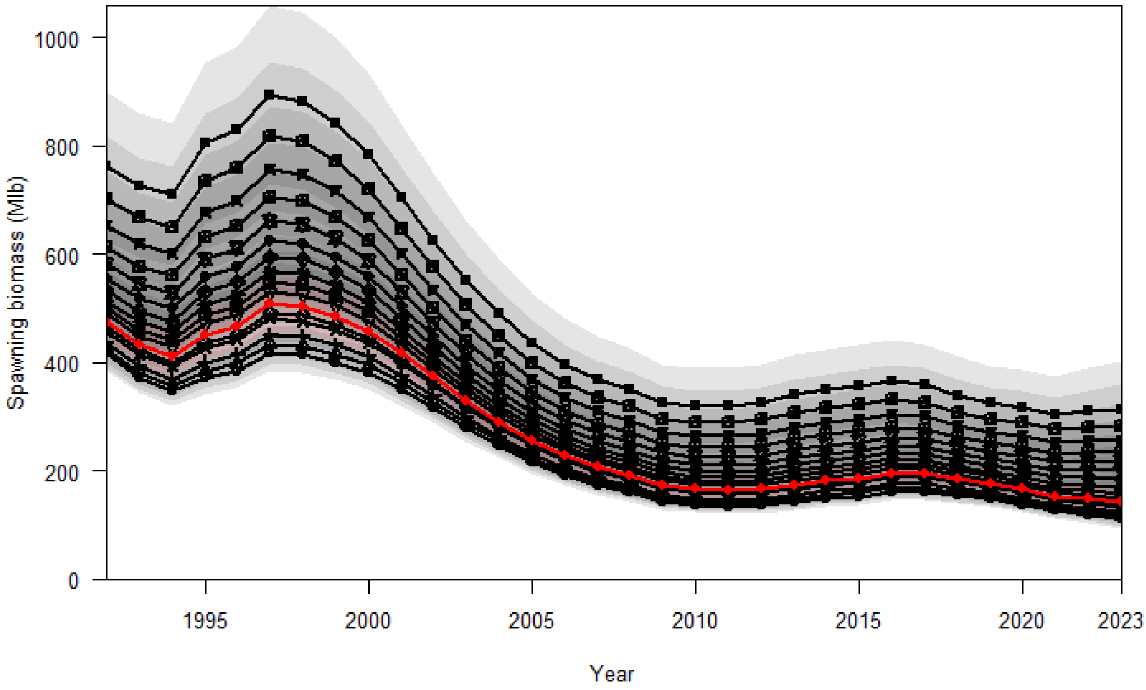


Figure 60. Spawning biomass estimates (lines and points) and corresponding 95% confidence intervals (shaded region) resulting from the likelihood profile on *M* from 0.10 to 0.25 for the coastwide short model. Red series denotes the fixed value used in the base case model.

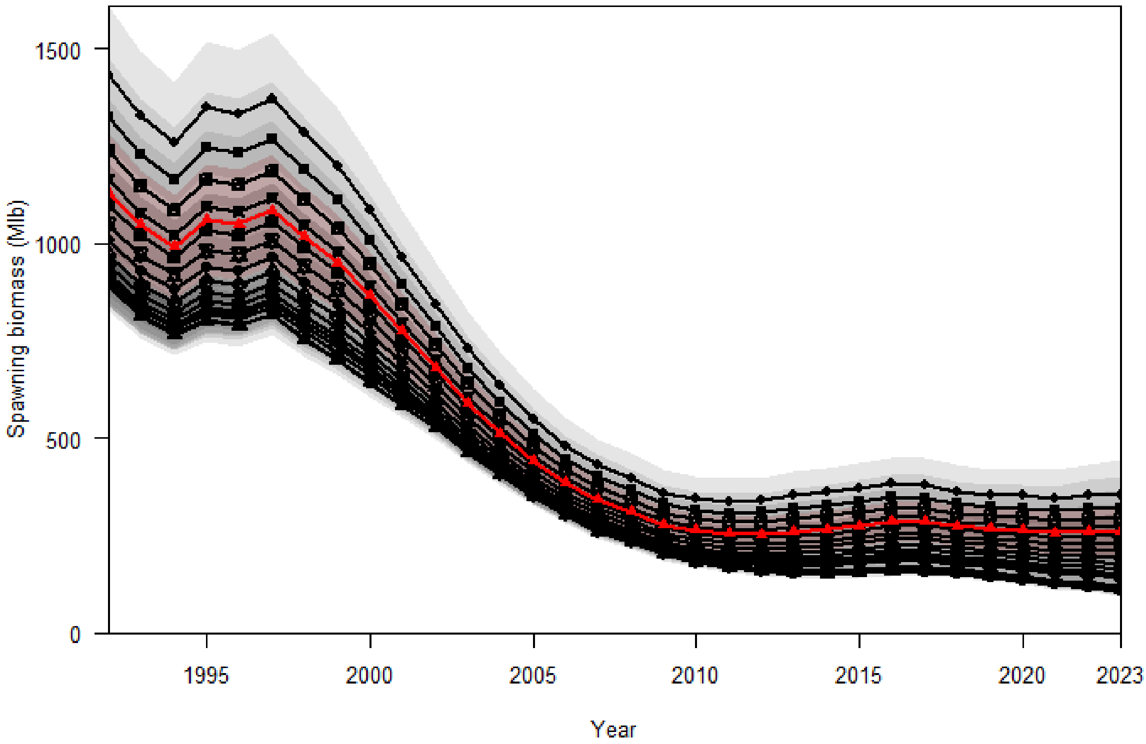


Figure 61. Spawning biomass estimates (lines and points) and corresponding 95% confidence intervals (shaded region) resulting from the likelihood profile on *M* from 0.10 to 0.25 for the AAF short model. Red series denotes the MLE (the base case model).

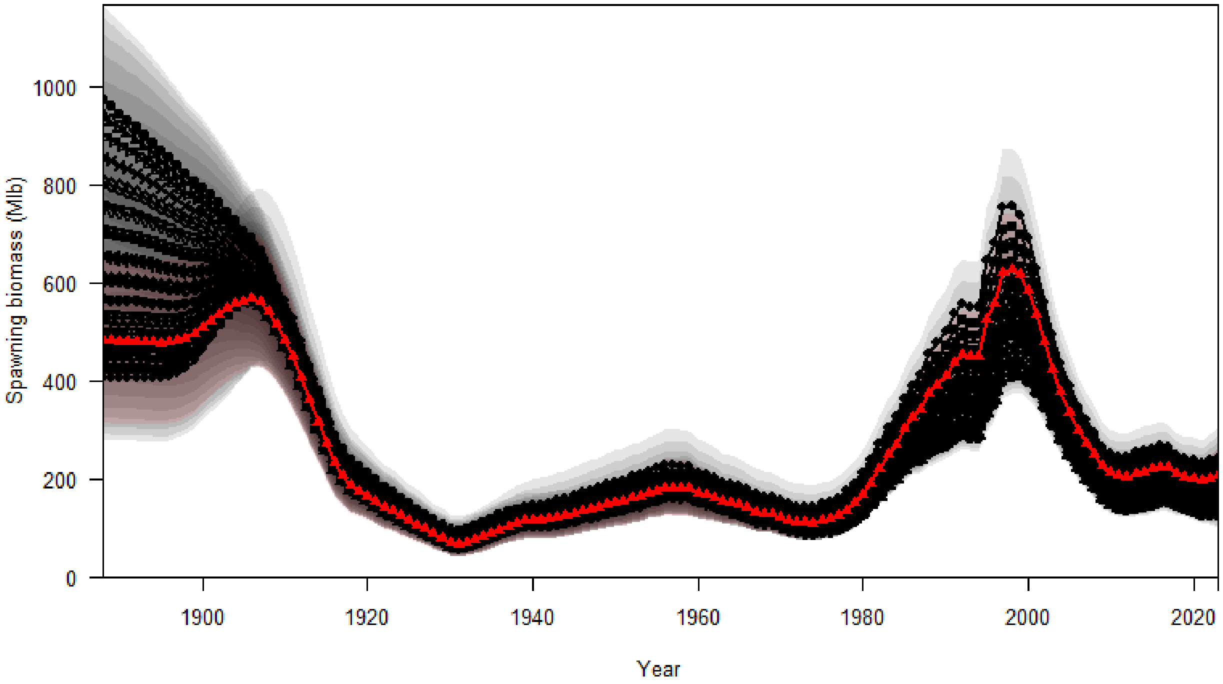


Figure 62. Spawning biomass estimates (lines and points) and corresponding 95% confidence intervals (shaded region) resulting from the likelihood profile on *M* from 0.10 to 0.25 for the coastwide long model. Red series denotes the MLE (the base case model).

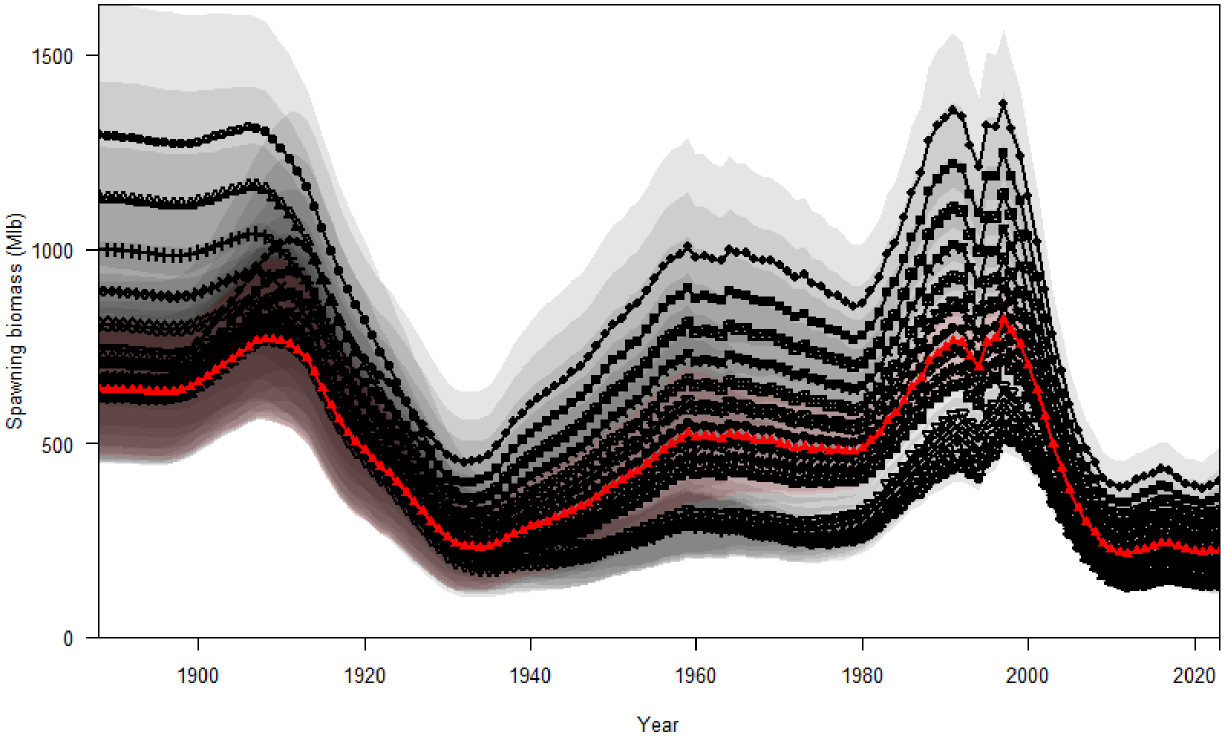


Figure 63. Spawning biomass estimates (lines and points) and corresponding 95% confidence intervals (shaded region) resulting from the likelihood profile on *M* from 0.10 to 0.25 for the AAF long model. Red series denotes the MLE (the base case model).

Retrospective analyses

The halibut model used from 2006 until 2011 was plagued by a very strong retrospective pattern, both in the scale of the most recent stock size estimates as well as the trend in those estimates (Stewart and Martell 2014; Stewart et al. 2013a). The solution to this problem was additional flexibility for process error (temporal variability) in the selectivity curves for both the fishery and survey representing not just gear (or ‘contact’) selectivity but also spatial availability.

Retrospective analyses were conducted for these preliminary 2022 models by sequentially removing the terminal four years of data from the model (a five-year retrospective, since the terminal year currently contains no information other than mortality projections). Limiting this approach to the most recent four years of data allows the models to be informed by at least one year of commercial fishery sex-ratio data, and therefore does not require a major change in assumptions within the retrospective (as was the case in the 2019 assessment; Stewart and Hicks 2019b).

All of the four models showed very little retrospective change as the terminal years of data were removed from the models (Figure 64-66). This an improvement over recent models which had modest trends and/or variability, although mostly confined to lie within annual confidence intervals. The cause of this reduced retrospective behavior appears to be the allowance for the scale of male selectivity to be time-varying. This effectively separates the most recent dynamics from the scaling of the fishery across all earlier years.

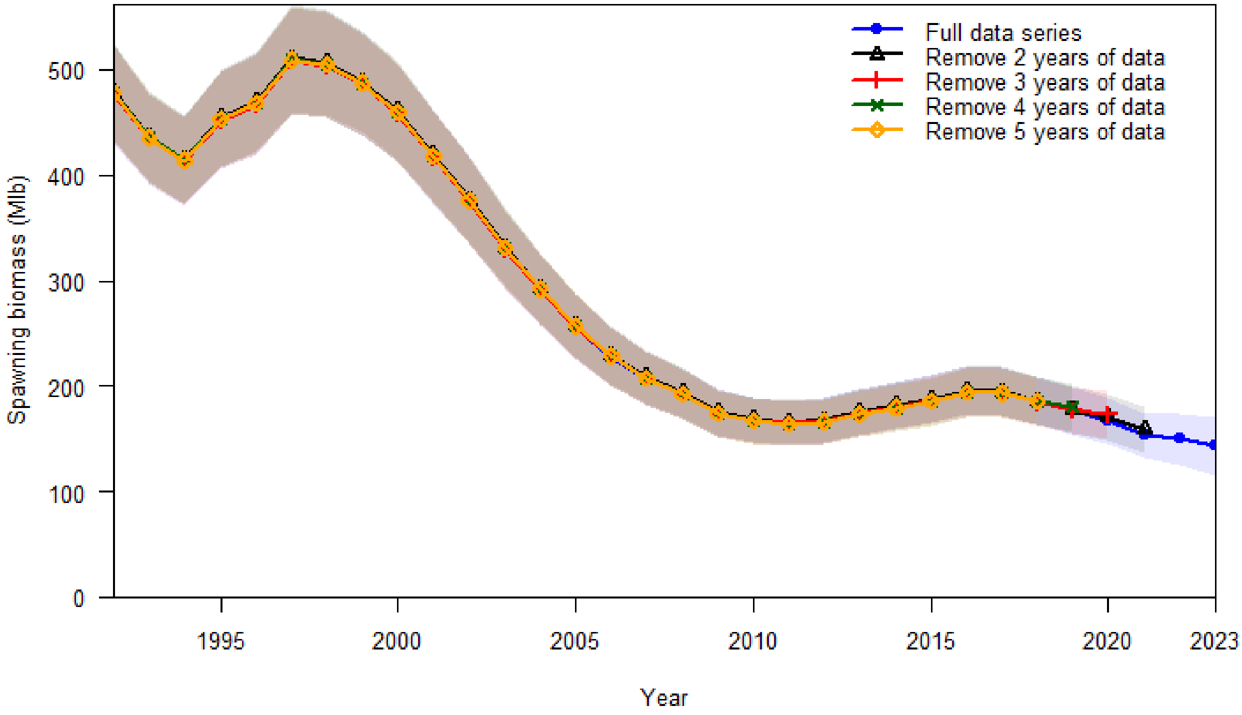


Figure 64. Five-year retrospective analysis of spawning biomass (1st year is a projection with no data) based on the coastwide short model.

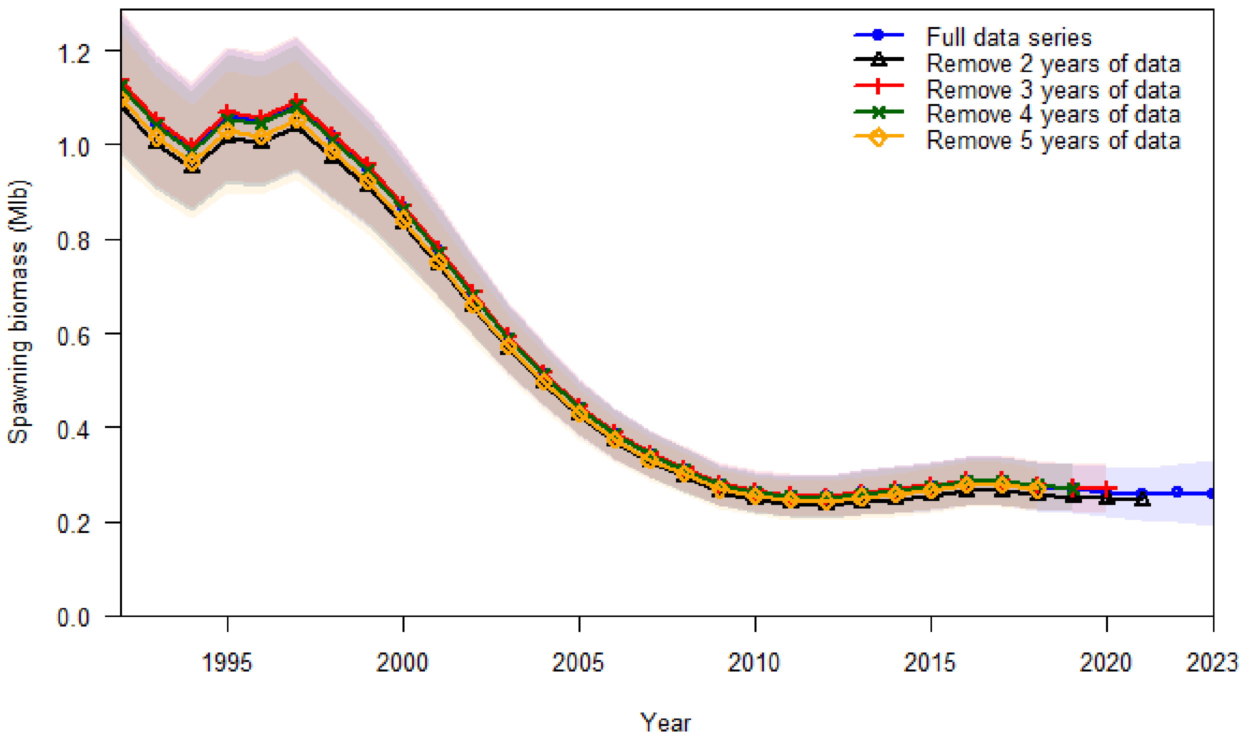


FIGURE 65. Five-year retrospective analysis of spawning biomass (1st year is a projection with no data) based on the AAF short model. Note that the y-axis is in billions of pounds.

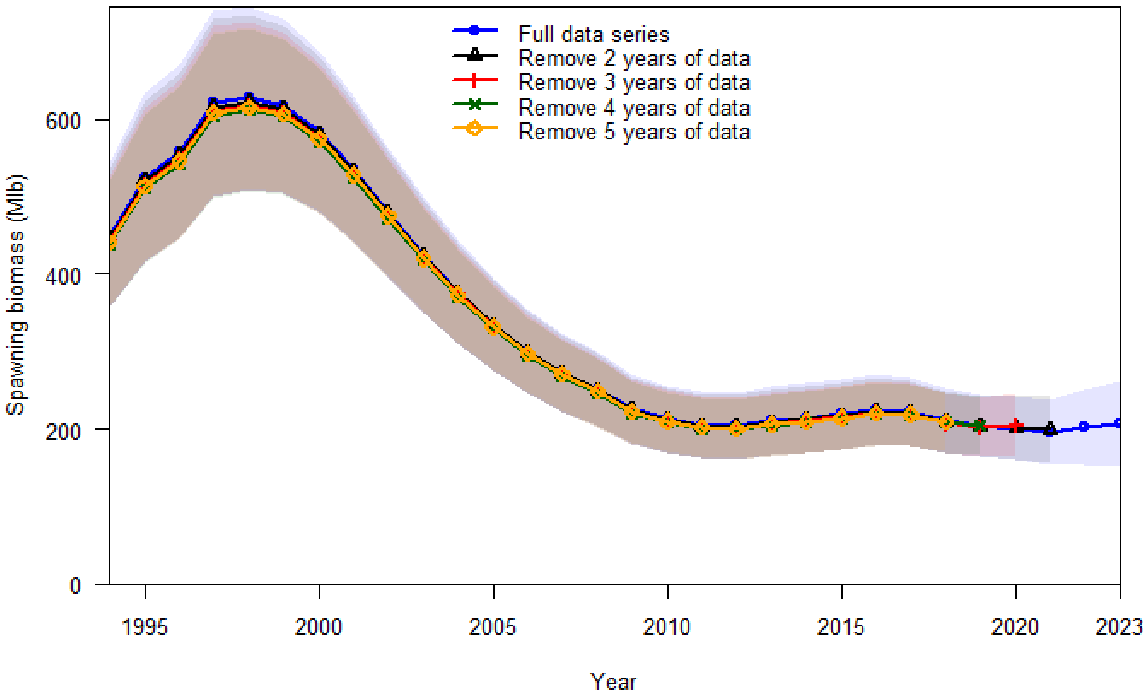
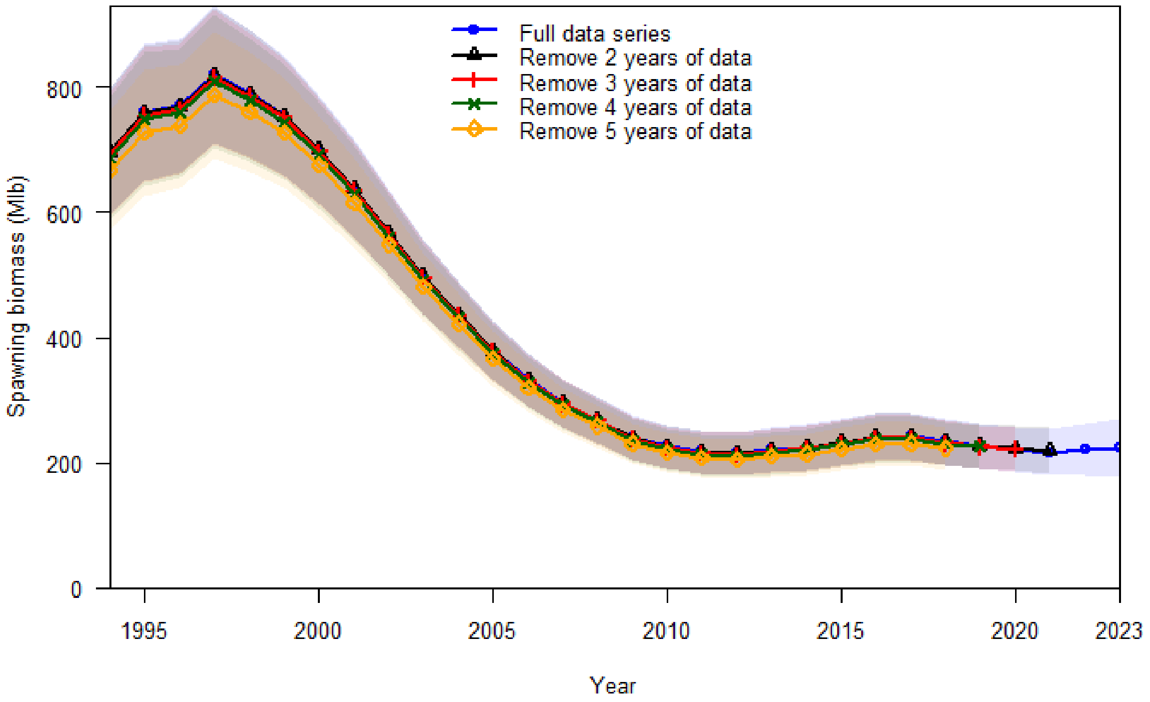


FIGURE 66. Five-year retrospective analysis of spawning biomass (1st year is a projection with no data) based on the coastwide long model. Time-series is truncated in 1994 so that differences in the terminal years are more visible.



Figures 67. Five-year retrospective analysis of spawning biomass (1st year is a projection with no data) based on the AAF long model. Time-series is truncated in 1994 so that differences in the terminal years are more visible.

Bayesian analysis

The 2019 stock assessment included a substantial evaluation of Bayesian integration for the short coastwide model (Stewart and Hicks 2019b). This effort did not produce substantially different results from the maximum likelihood and asymptotical variance methods (Fournier et al. 2012) routinely employed. However, there are a number of potential benefits to using an explicitly Bayesian approach, including better characterization of uncertainty (Magnusson et al. 2012) and a more directly interpretable characterization of the probability distributions. There is also the potential for differences in the results of Bayesian analyses due to the right-skewed nature of some distributions for key parameter and management-related quantities in complex fisheries models (Stewart et al. 2013b).

In aggregate, the 2019 results suggested that the asymptotic distributions were a reasonable approximation for the full posterior distributions in these models, and also that the process of regularizing the selectivity parameters and removing some deviations to improve integration did not having an appreciable effect on the solution. This is generally consistent with studies of process error where overparameterizing (adding the capability for variation when it wasn't present) was generally found to be unbiased, and therefore preferable to underparameterizing when temporal variability was present (e.g., Martell and Stewart 2014; Stewart and Monnahan 2017).

Additional Bayesian analysis was not included in this preliminary 2022 assessment. However, if a multi-year assessment approach was to become part of a future management procedure for the IPHC more time could be devoted to exploring Bayesian models.

Other uncertainty considerations

There are many important sources of uncertainty not captured in the four models included in this ensemble. These include myriad alternative structural assumptions such as spatially explicit population dynamics, connection with Russian waters, alternative stock-recruitment functions, time-varying mortality, different data weighting approaches, and many others. There are also several tractable sources of projection uncertainty that are not in the current approach, including uncertainty in projected weight-at-age (although the sensitivity of this was investigated at SRB request in 2016 and found to be low), projected selectivity, and projected fisheries mortality.

Within the modelled time-series there are also data-related uncertainties that could be addressed via a range of alternative approaches. Uncertainty in the time series of mortality for these models is not currently captured, as they are treated as inputs and assumed to be known without error. In previous assessments, sensitivity analyses have been conducted to the degree of discard mortality in the commercial fishery, potential effects of unobserved whale depredation, as well as to the magnitude of total bycatch mortality. In concept, these types of uncertainties could be explicitly included in the models; however, full estimation of catch in statistical catch-at-age models generally requires other stabilizing assumptions, so direct integration of this uncertainty may still prove challenging.

Additional sources of uncertainty and avenues for development are identified in the Research Priorities section below.

The ensemble

Model-integrated quantities are used as the primary output for stock assessment results, as well as the basis for decision table probabilities (Stewart and Hicks 2019a). All quantities of management interest are integrated for the recent time period (1992+), for which all four sets of model results are available. These quantities include: spawning biomass, relative spawning biomass, and the Spawning Potential Ratio (SPR; summarized as fishing intensity, $F_{XX\%}$, where the $XX\%$ represents SPR). Decision table quantities are divided into four categories: stock trend (which is the only set of metrics that are independent of any harvest strategy related assumptions), stock status, fishery trend, and fishery status. Integration is performed for all these quantities using the basic approach outlined below.

Methods

The basic approach to model integration remains unchanged from the 2015 and subsequent analyses. A sample of random draws is created from the output from each of the models included in the ensemble. For the spawning biomass time-series, the estimates and associated standard deviations for female spawning biomass from each of the four models were extracted from the report file. A vector of length n is created for each model (m), where the relative weight (w_m) is simply the relative fraction of the total draws across all models comprised by n_m:

$$w_m = \frac{n_m}{\sum_m n_m}$$

This approach allows for easily adjusted weighting of models. Routine reporting of results uses $\sum_m n_m$for all models equal to twenty million, this has been found to produce negligible Monte-Carlo error even in the tails of extremely skewed distributions, creating robust and stable reporting of all quantities of interest with smooth distribution. Although this choice could potentially be optimized for each statistic of interest, current integration code (in *R*) does not represent a constraining step in the analysis.

The harvest strategy employs a control rule that reduces the coastwide SPR target linearly from the interim 'reference level' at $SB_{30\%}$ to zero at $SB_{20\%}$. The calculation of relative spawning biomass was updated in the 2019 assessment to use a dynamic estimate of 'unfished' biomass calculated for each year of the time-series. This calculation replays the entire time-series, without the fishing mortality, assuming the same parameter values (including recruitment deviations) but accounting for the different level of spawning biomass projected for each year and its effect on subsequent expected (pre-deviation) recruitment in each year. At that time, the variance of this quantity and the covariance with estimated spawning biomass in each year was unavailable, so an approximation was developed (Stewart and Hicks 2019b). Subsequently, in 2020 the dynamic unfished biomass calculation was added to the derived quantities with variance calculations in stock synthesis, and so the approximation is no longer needed (Methot Jr et al. 2020a). This has been an important improvement as the covariance in estimated and unfished dynamic spawning biomass is an important contributor to the variance of the IPHC's reference points.

Evaluation of weighting based on predictive skill

Previous Pacific halibut assessments have applied equal weighting of all four models. However, weighting based on several potential approaches has been considered since the 2015 stock assessment (Stewart and Martell 2016). Briefly these have included:

> *AIC* – but this is known to be highly dependent on data weighting, and can only be applied in cases where the same data sets are being fit by all models under consideration
>
> *Strength of retrospective patterns* – perhaps relative to a 'null' distribution for a statistic like Mohn's rho (Mohn 1999) based on simulation (Hurtado-Ferro et al. 2015); while helpful to diagnose model performance, it does not necessarily indicate a 'good' model, as evidenced by the fact that a static prediction will have no retrospective pattern at all.
>
> *Fit to the FISS index* – without an AIC-type correction, there is no penalty for overparameterized models
>
> *Expert opinion* – this is subjective, and the tendency has been to revert to equal weighting in the absence of strong evidence to the contrary.

Meanwhile, exploration of model diagnostics for integrated models has highlighted other approaches to comparing model performance (Carvalho et al. 2021) and in particular the Mean

Absolute Standardized Error (MASE; Hyndman and Koehler 2006) may be particularly relevant for weighting stock assessment models:

$$MASE = \frac{\frac{1}{n}\sum_{t=1}^{n} |O_t - E_t|}{\frac{1}{n}\sum_{t=1}^{n} |O_t - O_{t-1}|}$$

Where *O* indicates the observation at time *t*, *E* the prediction (or expected value); calculations can be averaged over any number of years or lags relevant to the predictive problem. As defined, MASE estimates must be positive, and the range of values is interpreted as:

>1: model predictive skill is worse than the naïve prediction (last year's index) – model not worth pursuing further

1: model predictive skill is exactly equal to the naïve prediction

<1: model predictive skill exceeds that of the naïve prediction

0: model predictions perfectly match subsequent observations

This basic calculation available in the literature does not account for the observation error associated with each annual index. Conceptually, it does not make sense to treat lack of predictive skill for a year's index with a very large variance (some or all of the lack of skill may actually be observation error) equally with a year that is very precisely observed. We therefore extended the MASE calculation to use a standardized deviation rather than a raw deviation. This did not change the behavior or interpretation of the MASE values, the only addition being the standard deviation of the observation (σ_t) at time *t*:

$$MASE = \frac{\frac{1}{n}\sum_{t=1}^{n} \left|\frac{O_t - E_t}{\sigma_t}\right|}{\frac{1}{n}\sum_{t=1}^{n} \left|\frac{O_t - O_{t-1}}{\sigma_t}\right|}$$

This 'standardized' MASE statistic inherently accounts for over- or under-parameterization as it is concerned only with predictive skill. A major challenge to its widespread application is the need to determine which quantity (or quantities) should be used to evaluate predictive skill. In the case of Pacific halibut, this choice is simple: the FISS index closely tracks both the spawning biomass and the biomass available to the commercial fishery. Therefore, the relative trend in the FISS index will be directly indicative of the change in management quantities in the upcoming year. Second, the FISS index is also used as a step in the allocation of mortality limits, so the entire management procedure depends on its value each year.

For the appropriate time lag, a one-year ahead prediction is most relevant for Pacific halibut, since models are currently updated annually (although this could easily be modified for a management procedure with a two-year or longer lag between assessments). We might expect the predictive skill of each model to vary over time, and also the challenge of the prediction - years with very small changes from the previous year's index are 'harder' for models to exceed the naïve prediction than those with large changes. There must also be some variability in annual model performance that we may want to average over, specifically, we may not want to substantially down-weight a particular model due to a single poor prediction if it has generally been performing well. To explore model performance further we report results for MASE calculations spanning the most recent 1-4 years.

Since the coastwide FISS index is comprised of a composite of the spatially weighted indices from each Biological Region, it is possible to apply the same weighting to AAF model data and predictions (accounting for catchability) and thereby develop a predicted FISS index for all four of the individual models. These predictions can then be compared using the MASE statistic and weighted as described above.

In order to turn the MASE statistic into a model weight we need to specify the scale of the weighting and the behavior at the end-points. In this case, for model (*m*) within the set of models (*M*) we use the relative MASE:

$$MASE\ weight_m = \frac{1 - MASE_m}{\sum_{m=1}^{M} 1 - MASE_m}$$

This approach ensures that a model that does not outperform the naïve prediction (MASE >= 1) will get zero weight, and that a set of models all perfectly predicting the next observation will receive equal weights.

The most important prediction from the set of models is for the unobserved year (in this case 2022; Figure 68), and it has been helpful in the past to consider these predictions as part of the decision-making process. However, this prediction cannot be validated until after the decision-making process has occurred.

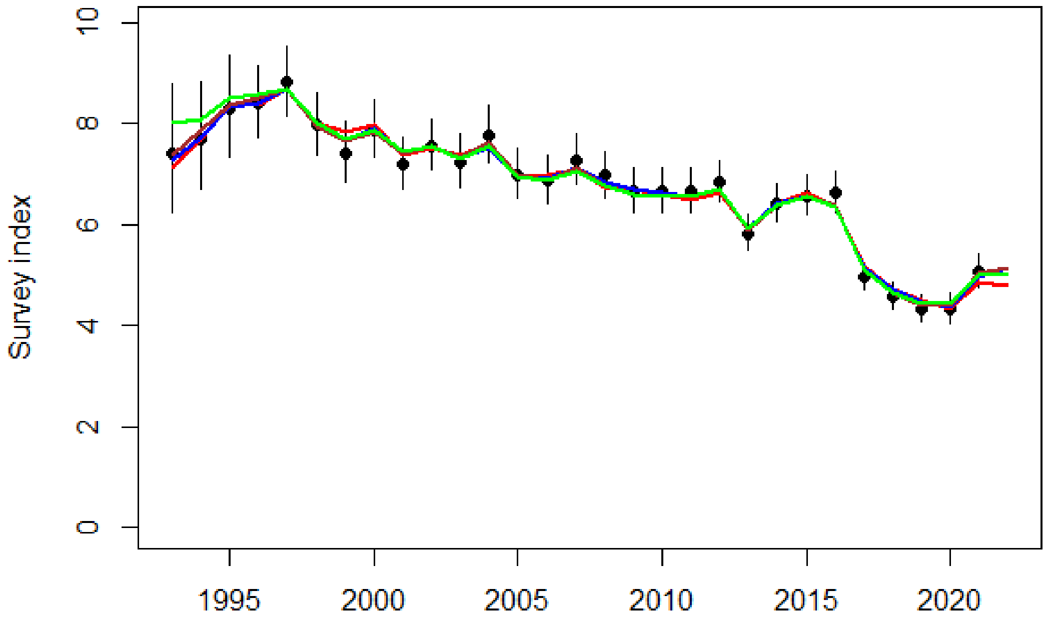


Figure 68. Predictions from each of the four models for the 2022 FISS observation using data through 2021 (black dots and CI).

In order to describe each model's predictive skill, a prediction was made for each recent year in the FISS time-series based on each step in a retrospective analysis. Specifically, one year of data was removed from the model fit, and then the prediction was made for the observed FISS index in the subsequent year. By working backwards within a single model, it is possible to evaluate how the predictions for each year's FISS index compared to the subsequent observation and the estimates from the model after the data had been included. Results for the coastwide short model are shown in Figure 69.

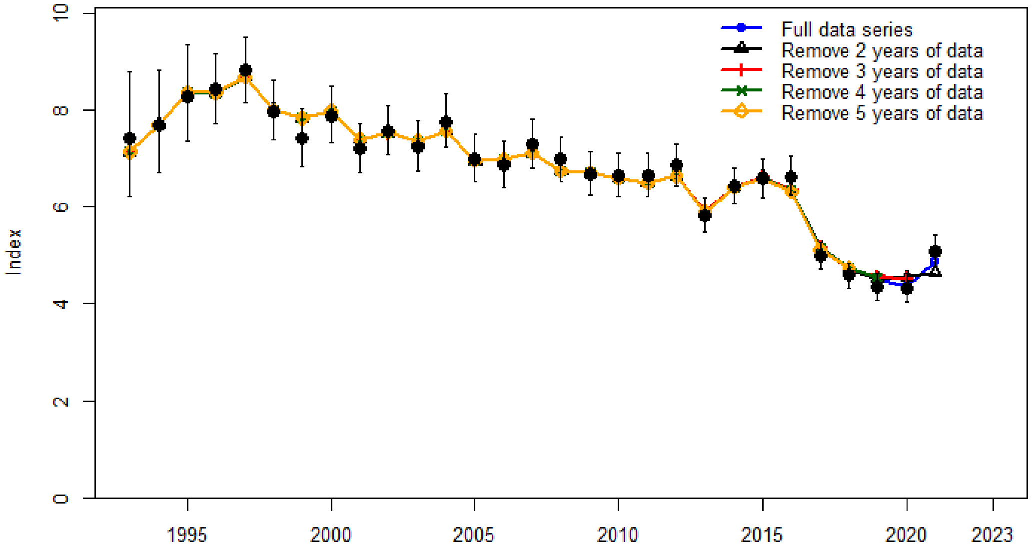


Figure 69. Predictions from the coastwide short model for the 2018-2021 FISS observations using data through 2017-2021 (black dots and CI).

When this process is repeated for all four individual models, the predictions can be visually compared at each step of the retrospective (Figure 70). The results indicate that the sharp increase in the 2021 FISS index was the most ‘challenging’ to predict (but may also be the most important recent test for these models) and that all models did appreciably better than the naïve prediction (the 2020 observed value). Further, because the 2020 observation was nearly equal to the 2019 observation, model predictions were similar to the naïve prediction. Comparison of the MASE scores averaging across the most recent 1, 2 , 3 and 4 years showed that all models performed better than the naïve prediction with MASE scores ranging from 0.44 – 0.94 (Table 19, Figure 71).

Converting these raw MASE scores into model weights, via the equation above, resulted in individual model weights varying from 9.3% to 38% across the range of models and years of averaging (Table 20, Figure 72). The aggregate ensemble results are relatively insensitive to weighting of the individual models, as the distributions are broadly overlapping and the weights are all similar. Specifically, the most extreme difference among model weights were for the three-year average MASE (9.3-38%) and the least extreme for the one-year MASE (20.5-28.3%; Table 20). Integrating over the full ensemble with these two vectors of weights produced quite similar spawning biomass trajectories (Figure 73-73). This is consistent with previous investigation of the effects of different weighting and new data on ensemble performance (Stewart and Martell 2015; Stewart and Hicks 2018).

This range of MASE weights does not clearly imply that one or more model’s contribution to the ensemble results should dramatically differ from the *status quo* assumption of equal weighting. However, there are several potential benefits to adopting a ‘dynamic’ or ‘self-weighting’ approach over static weights based on expert opinion. These include:

1) An objective basis for model weights based on predictive skill and logically tied to management information.
2) The ability to update weights each year (even during update assessments) based on the evolution of model predictive skill.

It might be expected that as stock dynamics change over time individual model skill in predicting upcoming management quantities would vary. The MASE calculation captures this evolution naturally and does not require an annual review and discussion of model weighting, except perhaps to ensure that the approach is performing as expected. Based on these benefits, we suggest that the 2022 stock assessment utilize MASE weights based on the most recent year (2022 for the final assessment, after the new data are available) of model prediction skill. Although potentially less stable than an average performance over recent years, weighting based on the terminal prediction will most closely represent the model skill if/when dynamics change over time.

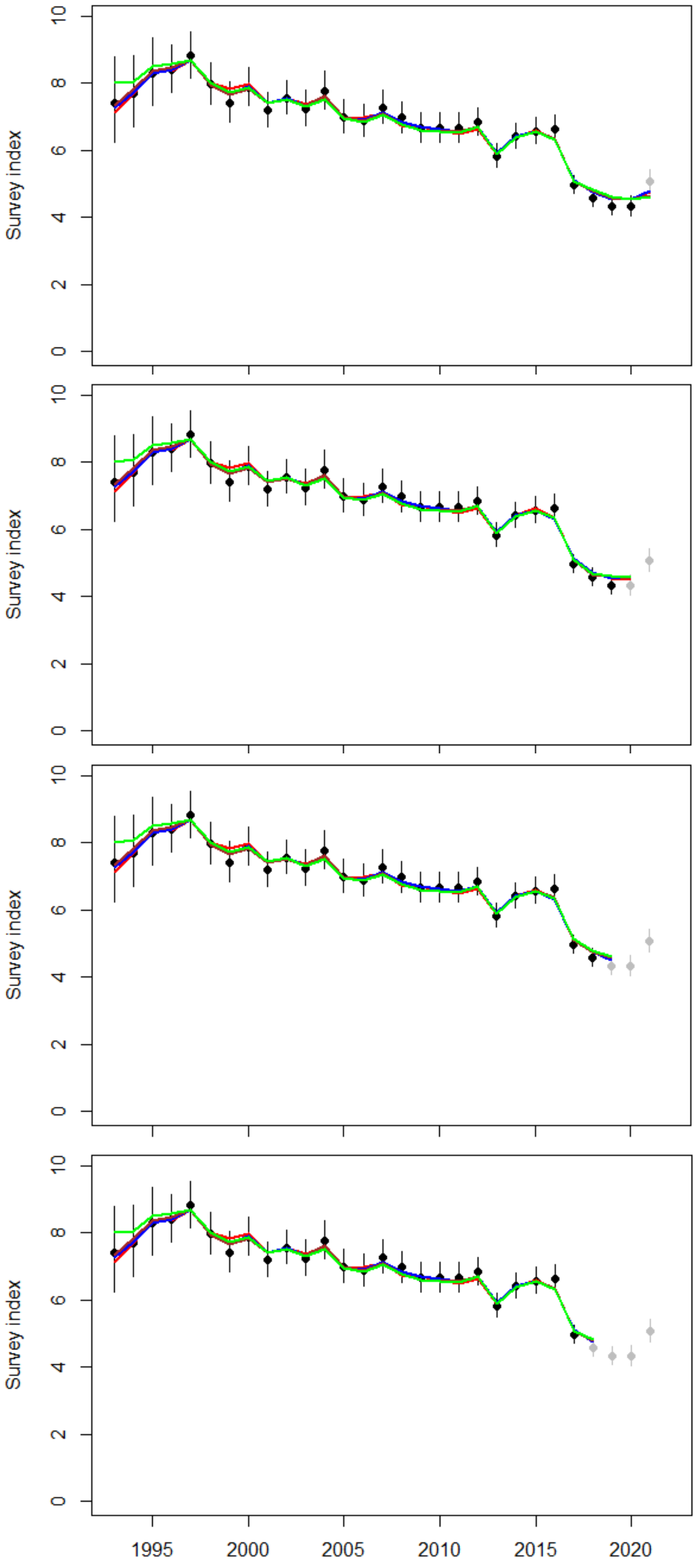


Figure 70. Predictions from each of the four models (colored lines) for the 2021 to 2018 (top to bottom panels) FISS observations (grey dots and CIs) using data through 2020 to 2017 (black dots and CI).

Table 19. One-year ahead standardized MASE estimates for each of the four stock assessment models averaged over the most recent 1, 2, 3, and 4 years.

	Model			
Years included	CW short	CW long	AAF short	AAF long
4	0.70	0.65	0.82	0.72
3	0.83	0.75	0.94	0.83
2	0.86	0.76	0.88	0.78
1	0.59	0.46	0.52	0.44

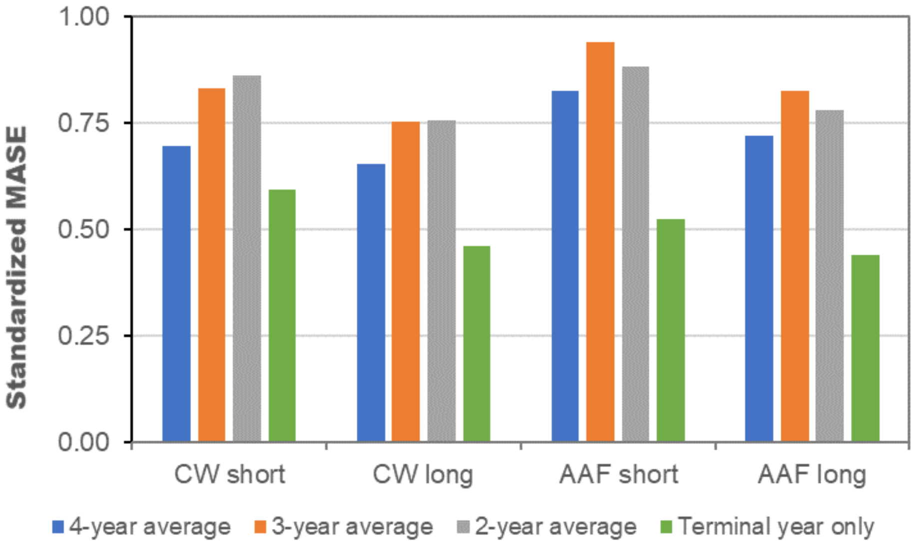


Figure 71. Comparison of standardized MASE estimates for each of the four models averaged over the most recent 1-4 years.

Table 20. One-year ahead standardized MASE weights for each of the four stock assessment models averaged over the most recent 1, 2, 3, and 4 years.

	Model			
Years included	CW short	CW long	AAF short	AAF long
4	27.5%	31.3%	15.8%	25.4%
3	26.0%	38.0%	9.3%	26.8%
2	19.1%	33.9%	16.4%	30.6%
1	20.5%	27.2%	24.0%	28.3%
Status quo weights	25.0%	25.0%	25.0%	25.0%

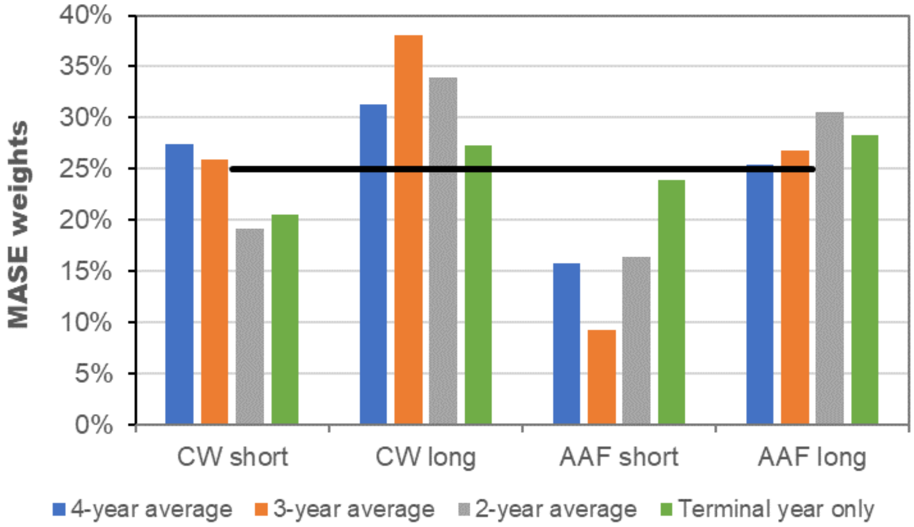


Figure 72. Comparison of standardized MASE weights for each of the four models averaged over the most recent 1-4 years. Horizontal line indicates the *status quo* equal weighting (25%).

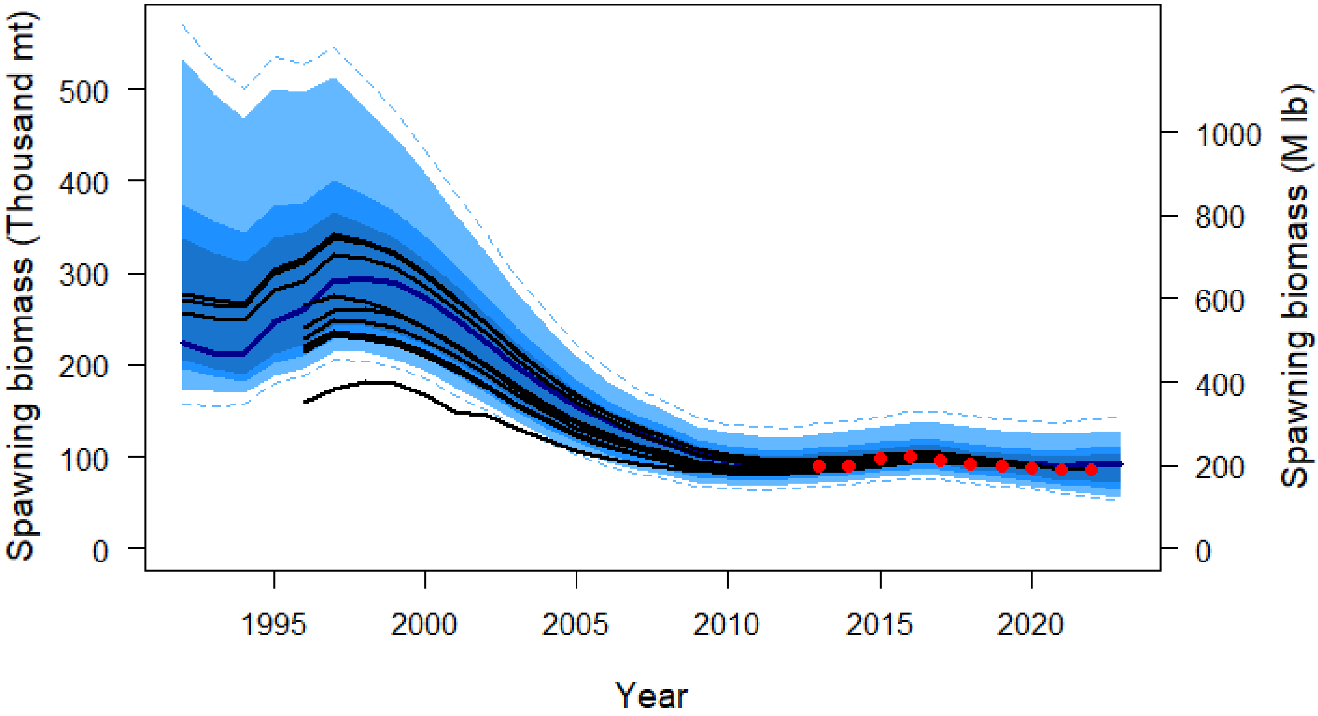


Figure 73. Comparison of the preliminary 2022 ensemble spawning biomass distribution based on the average MASE over the most recent three years (blue shading) to previous stock assessments (2012-2021; black lines, terminal estimates indicted by red dots).

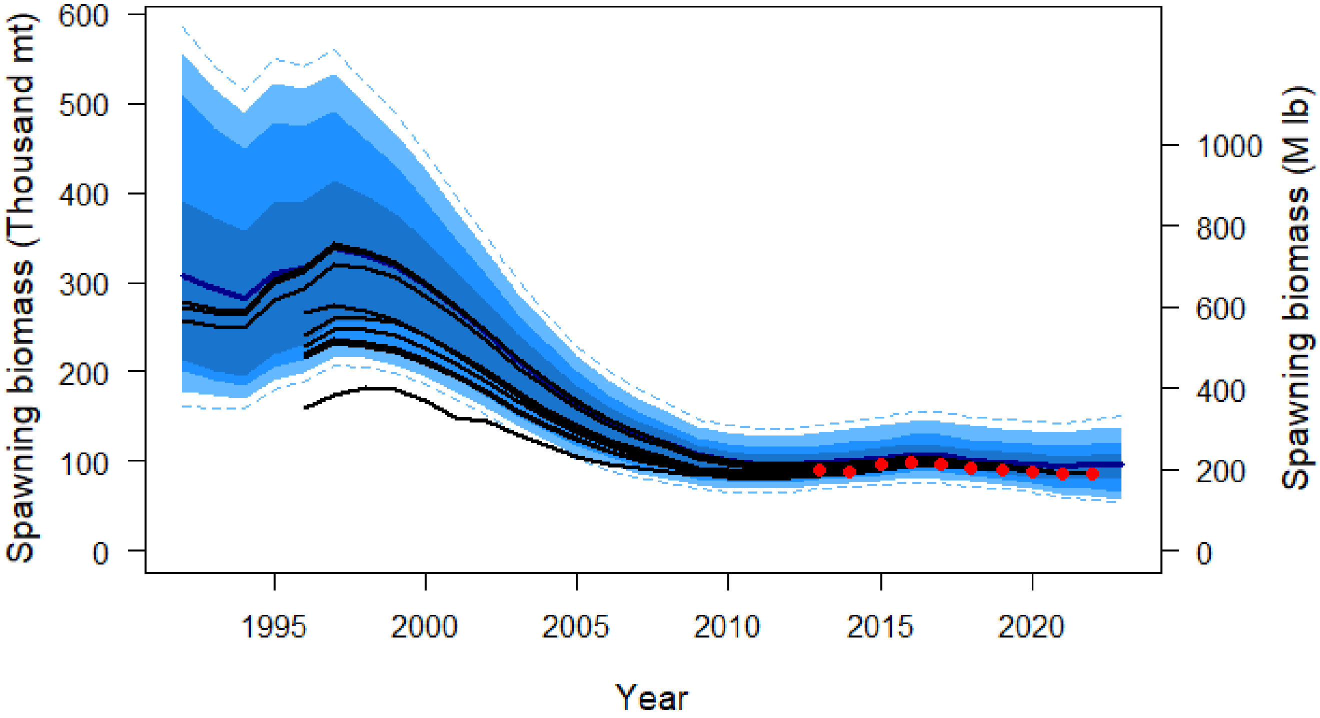


Figure 74. Comparison of the preliminary 2022 ensemble spawning biomass distribution based on terminal-year MASE (blue shading) to previous stock assessments (2012-2021; black lines, terminal estimates indicted by red dots).

Preliminary results for 2022

Comparison of the spawning biomass estimates from the four stock assessment models comprising the ensemble shows that the 95% intervals from any single model are substantially narrower than the aggregate (Figure 75). All four models indicate a similar overall trajectory, including the small increase in biomass over 2011-2016 and subsequent decrease as the effects of reduced recruitment during 2006-2011 (Figure 76; upper panel) graduate through to the spawning biomass. The differences in *M* among the four models suggest large absolute differences in recruitment estimates, but when scaled relative to the mean it is very clear that the estimates of relative strong and weak year classes are in close agreement (Figure 76; lower panel).

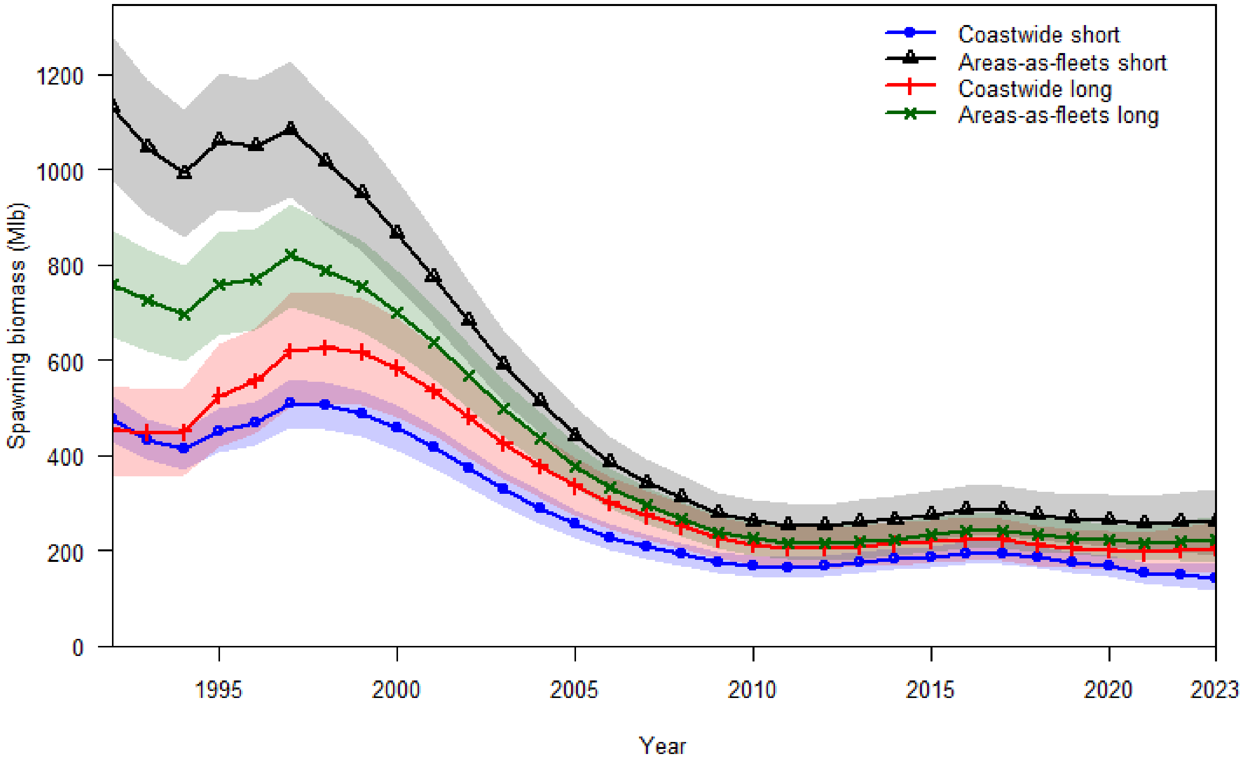


Figure 75. Comparison of spawning biomass time series (shaded regions indicate asymptotic approximations to the 95% confidence interval) from each of the preliminary models contributing to the 2022 ensemble.

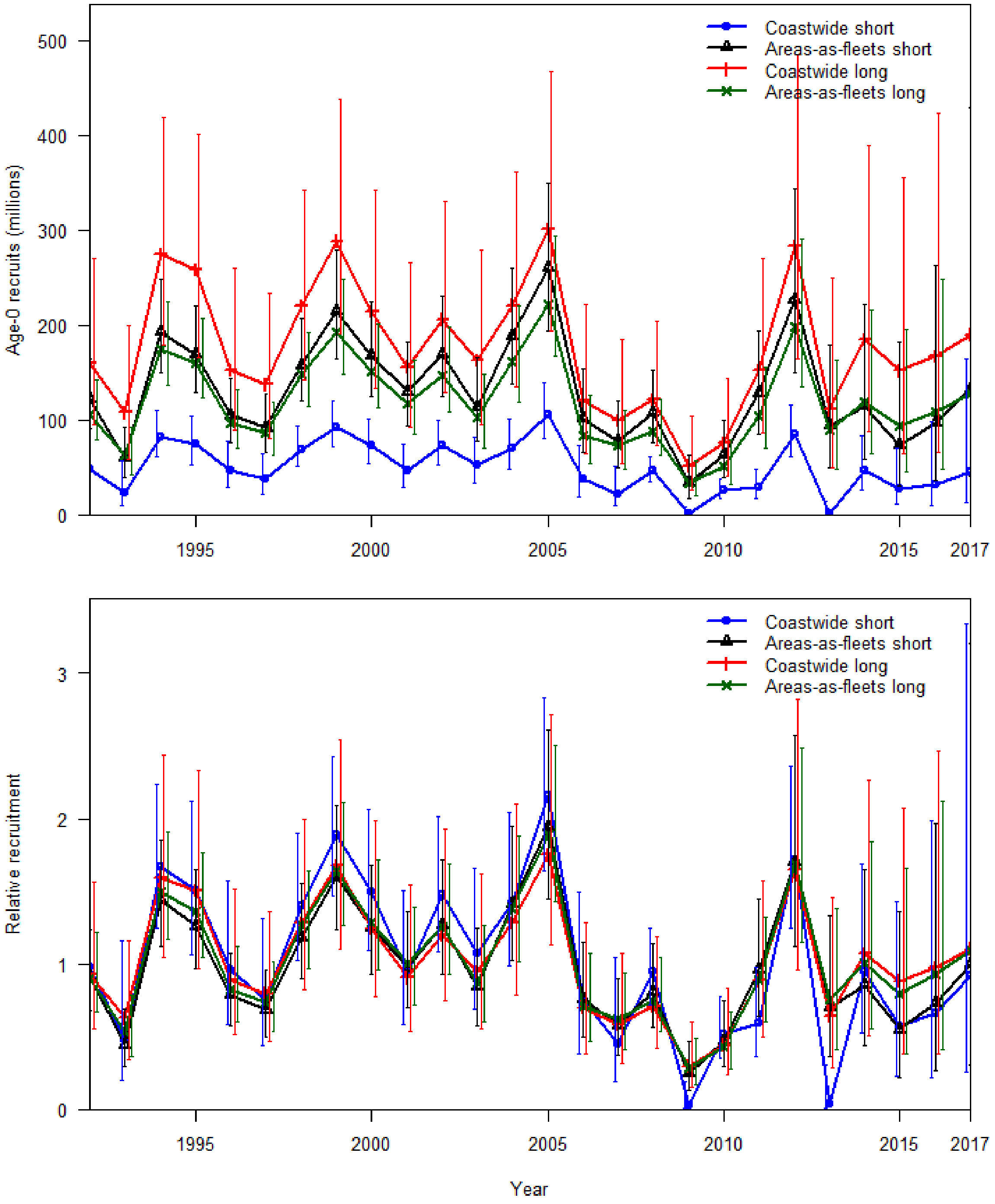


Figure 76. Comparison of recruitment time series (upper panel; vertical lines indicate asymptotic approximations to the 95% confidence interval) and relative recruitment series (each standardized to its mean; lower panel) from each of the preliminary models contributing to the 2022 ensemble.

Future development

Several extensions to this preliminary assessment will be possible for the final 2022 analysis. These include:

- Responses to suggestions and comments generated from SRB020 and SRB021.
- Addition of all 2022 data, extending existing time series (mortality, indices, ages, etc.).
- The sex-ratio of the 2021 commercial fisheries landings based on the IPHC's genetic assay will be completed and included in the final 2022 assessment.

In addition to the list of research priorities (longer list below), there are several potential avenues for development within and among the four models included in the ensemble.

The bootstrapping performed for this assessment provides a strong basis for objective interannual and among fleet weighting. Future work can now again focus on the likelihoods used for data weighting in this assessment. Options for compositional likelihoods, including those already evaluated to some degree for this assessment over the last several years (e.g., the Dirichlet-multinomial, logistic normal) continue to expand. A new candidate that can allow for automatic scaling and an estimated relationship between the observed proportion and the variance, the Tweedie distribution, is currently in press (J. Thorson, personal communication). Further, work on a calculation of composition residuals that improves upon the standard Pearson residuals currently employed by most stock assessments is also in preparation; these PIT residuals are more computationally intensive, but may have much improved distributional characteristics (Warton et al. 2017).

Other avenues for development include changes to the ensemble approach itself. The 2019 assessment explored expanding the number of models included in the ensemble to better capture the uncertainty in *M* that was missed through using a fixed value in the two short time-series models. By estimating *M* for the short AAF model in this assessment the integration of uncertainty is improved. The question of how to better address *M* in the short coastwide model remains. The next full assessment may need to explore whether structural changes could make *M* estimable and/or whether the fixed value of 0.15 is still appropriate given the increasing weight of evidence that *M* for Pacific halibut is higher, even after accounting for elevated *M* at the youngest ages.

As ensemble changes are evaluated, both weighting and technical efficiency should be considered. Technical costs of adding additional models to the ensemble include additional time spent running these additional models rather than exploring other sensitivities and identifying clear effects of newly available data during the very short assessment analysis period each fall. Pragmatically, there may be relatively little to be gained from increasing the ensemble in this manner beyond slightly smoother integrated distributions. As the IPHC's management procedure evolves, to potentially include multi-year assessments, there may be additional latitude for increased model and ensemble complexity.

The current ensemble is based on maximum likelihood estimates and asymptotic approximations to the posterior distributions for model parameters and derived quantities.

Bayesian posteriors represent a conceptually more appealing basis for probability distributions, and could better capture the full range and potential asymmetries in the distributions for model quantities (Magnusson et al. 2012; Stewart et al. 2013b). Bayesian integration may also allow for statistically correct treatment of variance parameters (such as the sigmas governing recruitment variability and selectivity or catchability process error) in the absence of true random effects capability in AD Model builder. Although it would be technically preferable to regularize and run all four assessment models as Bayesian analyses, at present this is technically infeasible given the tight time-line between data availability and the deadline for the annual stock assessment. The analysis time difference between minimization and full posterior integration, even using the most efficient methods available for the coastwide short model (see section above), is still too large. However, if the IPHC were to move to a more formal management procedure and/or to a multi-year mortality limit-setting process, the stock assessment could be conducted at a pace that would allow much greater reliance on Bayesian models.

Research priorities

The development of the IPHC's research priorities has been closely tied to the needs of the stock assessment and harvest strategy policy analyses, such that the IPHC's research projects will provide data, and hopefully knowledge, about key biological and ecosystem processes that can then be incorporated directly into analyses supporting the management of Pacific halibut. Research priorities for the Pacific halibut stock assessment can be delineated into three broad categories: improvements in basic biological understanding (including fishery dynamics), investigation of existing data series and collection of new information, and technical development of models and modelling approaches. The highest priority items in each of these categories are highlighted in the 5-year research plan and are expected to be the primary focus of ongoing efforts. However, it is helpful to maintain a longer list of items to inform future prioritization, to create a record of data and research needs, and to foster opportunistic and/or collaborative work on these topics when possible.

Biological understanding and fishery yield

Key areas for improvement in biological understanding include:

- *Highest priority*: Updating the current functional maturity schedule for Pacific halibut, including fecundity-weight relationships and the presence and/or rate of skip spawning.
- *Highest priority*: The stock structure of the Pacific halibut population. Specifically, whether any geographical components (e.g., Biological Region 4B) are isolated to a degree that modelling approximations would be improved by treating those components separately in the demographic equations and management decision-making process.
- *Highest priority*: Movement rates among Biological Regions at the adult, juvenile and larval stages remain uncertain and likely variable over time. Long-term research to inform these rates could lead to a spatially explicit stock assessment model for future inclusion into the ensemble.
- *Highest priority*: Improved understanding of discard mortality rates and the factors contributing to them may reduce potential biases in mortality estimates used for stock

assessment and allow for future reduction in mortality through improved handling practices

- The relative role of potential factors underlying changes in size-at-age is not currently understood. Delineating between competition, density dependence, environmental effects, size-selective fishing and other factors could allow improved prediction of size-at-age under future conditions.
- Improved understanding of recruitment processes and larval dynamics could lead to covariates explaining more or the residual variability about the stock-recruit relationship than is currently accounted for via the binary indicator used for the Pacific Decadal Oscillation.

Data related research

This section represents a list of potential projects relating specifically to existing and new data sources that could benefit the Pacific halibut stock assessment.

- *Highest priority*: Continued collection of sex-ratio from the commercial landings will provide valuable information for determining relative selectivity of males and females, and therefore the scale of the estimated spawning biomass, and the level of fishing intensity as measured by SPR. Potential methods for estimating historical sex-ratios from archived scales, otoliths or other samples should be pursued if possible.
- *Highest priority*: Evaluation of the magnitude of marine mammal depredation and tools to reduce it.
- The work of Monnahan and Stewart (2015) modelling commercial fishery catch rates could be used to provide a standardized fishery index for the recent time-series that would be analogous to the space-time model used for the FISS.
- A revised hook spacing relationship (Monnahan and Stewart 2017) could be included into IPHC database processing algorithms.
- There is a vast quantity of archived historical data that is currently inaccessible until organized, electronically entered, and formatted into the IPHC's database with appropriate meta-data. Information on historical fishery landings, effort, and age samples would provide a much clearer (and more reproducible) perception of the historical period.
- Additional efforts could be made to reconstruct estimates of subsistence harvest prior to 1991.
- Discard mortality estimates for the IPHC Regulatory Area 2B recreational fishery are currently unavailable, but there is an estimation system in place. Further work to develop these estimates would be preferable to the use of proxy rates from IPHC Regulatory Area 2C.
- NMFS observer data from the directed Pacific halibut fleet in Alaska could be evaluated for use in updating DMRs and the age-distributions for discard mortality. This may be more feasible if observer coverage is increased and if smaller vessels (< 40 feet LOA, 12.2 m) are observed in the future. Post-stratification and investigation of observed vs. unobserved fishing behavior may be required.

- Historical bycatch length frequencies and mortality estimates need to be reanalyzed accounting for sampling rates in target fisheries and evaluating data quality over the historical period.
- There are currently no comprehensive variance estimates for the sources of mortality used in the assessment models. In some cases, variance due to sampling and perhaps even non-sampling sources could be quantified and used as inputs to the models via scaling parameters or even alternative models in the ensemble.
- A space-time model could be used to calculate weighted FISS age-composition data. This might alleviate some of the lack of fit to existing data sets that is occurring not because of model misspecification but because of incomplete spatial coverage in the annual FISS sampling which is accounted for in the generation of the index, but not in the standardization of the composition information.

Technical development

There are a variety of technical explorations and improvements that could benefit the stock assessment models and ensemble framework. Although larger changes, such as the new data sets and refinements to the models presented in this document, naturally fit into the period full assessment analyses, incremental changes may be possible during updated assessments when and if new data or methods become available. Specifically, development is intended to occur in time for initial SRB review (generally in June), with only refinements made for final review (October), such that untested approaches are not being implemented during the annual stock assessment itself. Technical research priorities include:

- *Highest priority*: Maintaining consistency and coordination between MSE, and stock assessment data, modelling and methodology.
- *Highest priority*: ‘Leading’ parameter estimation. Building on the improvements to estimation of *M* in the short AAF model in this assessment, focus should be on estimation of *M* in the short coastwide model.
- *Highest priority*: Evaluation of estimating (Thorson 2019) rather than tuning (Francis 2011; Francis 2016) the level of observation and process error in order to achieve internal consistency and better propagate uncertainty within each individual assessment model. This could include tools like the 2d-autoregressive smoother for selectivity, the Dirichlet multinomial, Tweedie, and other features now implemented or in development in stock synthesis.
- Continued refinement of the ensemble of models used in the stock assessment. This may include investigation of alternative approaches to modelling selectivity that would reduce relative down-weighting of certain data sources (see section above), evaluation of additional axis of uncertainty (e.g., steepness, as explored above), or others.
- Exploration of methods for better including uncertainty in directed and non-directed discard mortalities in the assessment (now evaluated only via alternative mortality

projection tables or model sensitivity tests) in order to better include these sources uncertainty in the decision table. These could include explicit discard/retention relationships, including uncertainty in discard mortality rates, and allow for some uncertainty directly in the magnitude of mortality for these sources.

- Bayesian methods for fully integrating parameter uncertainty may provide improved uncertainty estimates within the models contributing to the assessment, and a more natural approach for combining the individual models in the ensemble (see section above).
- Alternative model structures, including a growth-explicit statistical catch-at-age approach and a spatially explicit approach may provide avenues for future exploration. Efforts to develop these approaches thus far have been challenging due to the technical complexity and data requirements of both. Previous reviews have indicated that such efforts may be more tractable in the context of operating models for the MSE, where conditioning to historical data may be much more easily achieved than fully fitting an assessment model to all data sources for use in tactical management decision making.

Acknowledgements

IPHC datasets comprise a wide array of sources based on extensive sampling and reporting efforts by state and national agencies in the U.S. and Canada. The IPHC’s annual stock assessment benefits from the hard work of all of its current and former employees providing high-quality data sets as comprehensive as any used for fisheries analysis. The Scientific Review Board and national science advisors have provided extensive guidance and constructive criticism of the treatment of data sources, the individual models and the stock assessment ensemble. Ray Webster leads, or contributes to, many of the supporting data analyses on which the assessment is based.

References

Adams, G.D., Holsman, K.K., Barbeaux, S.J., Dorn, M.W., Ianelli, J.N., Spies, I., Stewart, I.J., and Punt, A.E. 2022. An ensemble approach to understand predation mortality for groundfish in the Gulf of Alaska. Fisheries Research **251**. doi:10.1016/j.fishres.2022.106303.

Bryan, M., and Ferriss, B. 2021. Assessment of the shallow-water flatfish stock complex in the Gulf of Alaska. Chapter 4. Gulf of Alaska SAFE. 49 p.

Bryan, M.D., and Palsson, W. 2021. Assessment of Northern and Southern rock sole (*Lepidopsetta polyxstra* and *bilineata*) stocks in the Gulf of Alaska. Chapter 4. NPFMC Gulf of Alaska SAFE. 150 p.

Bryan, M.D., Shotwell, K., Zador, S., and Ianelli, J. 2020a. Assessment of the Kamchatka flounder stock in the Bering Sea and Aleutian Islands. Chapter 7. NPFMC Bering Sea and Aleutian Islands SAFE. 91 p.

Bryan, M.D., Barbeaux, S.J., Ianelli, J., Zador, S., and Hoff, J. 2020b. Assessment of the Greenland turbot stock in the Bering Sea and Aleutian Islands. Chapter 5. NPFMC Bering Sea and Aleutian Islands SAFE. 96 p.

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

Clark, W.G. 2003. A model for the world: 80 years of model development and application at the international Pacific halibut commission. Natural Resource Modeling **16**(4): 491-503.

Clark, W.G. 2004a. Nonparametric estimates of age misclassification from paired readings. Canadian Journal of Fisheries and Aquatic Sciences **61**: 1881-1889.

Clark, W.G. 2004b. Statistical distribution of IPHC age readings. IPHC Report of Assessment and Research Activities 2003. p. 99-110.

Clark, W.G., and Hare, S.R. 2002a. Effects of climate and stock size on recruitment and growth of Pacific halibut. North American Journal of Fisheries Management **22**: 852-862.

Clark, W.G., and Hare, S.R. 2002b. Effects of climate, stock size, and region on recruitment and growth of Pacific halibut. IPHC Report of Assessment and Research Activities 2001. p. 169-190.

Clark, W.G., and Hare, S.R. 2006. Assessment and management of Pacific halibut: data, methods, and policy. International Pacific Halibut Commission Scientific Report No. 83, Seattle, Washington. 104 p.

Cox, S.P., Ianelli, J., and Mangel, M. 2014. IPHC Scientific Review Board Meeting. June 2014., 3 p.

Cox, S.P., Ianelli, J., and Mangel, M. 2017. 4.7 Reports of the IPHC Scientific Review Board (SRB). IPHC Report of Assessment and Research Activities 2016. p. 443-452.

Dahlheim, M.E., Cahalan, J., and Breiwick, J.M. 2022. Interactions, injuries, and mortalities of killer whales (Orcinus orca) observed during fishing operations in Alaska. Fishery Bulletin **120**(1): 79-94. doi:10.7755/fb.120.1.8.

Du, J. 2014. Uncertainty and ensemble forecast. National Weather service. 42 p.

Forsberg, J.E., and Stewart, I.J. 2015. Re-ageing of archived otoliths from the 1920s to the 1990s. IPHC Report of Assessment and Research Activities 2014. p. 405-428.

Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., and Sibert, J. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software **27**(2): 233-249.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences **68**: 1124-1138.

Francis, R.I.C.C. 2016. Revisiting data weighting in fisheries stock assessment models. Fish. Res. doi:10.1016/j.fishres.2016.06.006.

Goethel, D.R., and Berger, A.M. 2017. Accounting for spatial complexities in the calculation of biological reference points: effects of misdiagnosing population structure for stock status indicators. Canadian Journal of Fisheries and Aquatic Sciences **74**(11): 1878-1894. doi:10.1139/cjfas-2016-0290.

Goethel, D.R., Hanselman, D.H., Rodgveller, C.J., Echave, K.B., Williams, B.C., Shotwell, S.K., Sullivan, J.Y., Hulson, P.F., Malecha, P.W., Siwicke, K.A., and Lunsford, C.R. 2021. Assessment of the sablefish stock in Alaska. NPFMC Bering Sea, Aleutian Islands and Gulf of Alaska SAFE. Chapter 3. 347 p.

Grandin, C.J., and Forrest, R.R. 2017. Arrowtooth flounder (Atheresthes stomias) stock assessment for the west coast of British Columbia. DFO Can. Sci. Res. Doc. 2017/025. v + 87 p.

Hamel, O.S. 2014. A method for calculating a meta-analytical prior for the natural mortality rate using multiple life history correlates. ICES Journal of Marine Science **72**(1): 62-69. doi:10.1093/icesjms/fsu131.

Hamill, T.M., Brennan, M.J., Brown, B., DeMaria, M., Rappaport, E.N., and Toth, Z. 2012. NOAA's Future Ensemble-Based Hurricane Forecast Products. Bulletin of the American Meteorological Society **93**: 209-220.

He, X., Pearson, D.E., Field, J.C., Lefebvre, L., and Key, M. 2013. Status of the U.S. pacific sanddab resource in 2013. Pacific Fishery Management Council. 344 p.

Hicks, A., and Stewart, I. 2022. Update on the IPHC secretariat MSE program of work (2021-2023). IPHC-2022-AM098-12. 16 p.

Holt, K.R., Starr, P.J., Haigh, R., and Krishka, B. 2016. Stock assessment and harvest advice for rock sole (*Lepidopsetta spp.*) in British Columbia. DFO Can. Sci. Advis. Sec. Res. Doc. 2016/009. ix + 256 p.

Hurtado-Ferro, F., Szuwalski, C.S., Valero, J.L., Anderson, S.C., Cunningham, C.J., Johnson, K.F., Licandeo, R., McGilliard, C.R., Monnahan, C.C., Muradian, M.L., Ono, K., Vert-Pre, K.A., Whitten, A.R., and Punt, A.E. 2015. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. ICES Journal of Marine Science **72**(1): 99-110. doi:10.1093/icesjms/fsu198.

Hyndman, R.J., and Koehler, A.B. 2006. Another look at measures of forecast accuracy. International Journal of Forecasting **22**(4): 679-688. doi:10.1016/j.ijforecast.2006.03.001.

Ianelli, J., Holsman, K.K., Punt, A.E., and Aydin, K. 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. Deep Sea Research Part II: Topical Studies in Oceanography **134**: 379-389. doi:10.1016/j.dsr2.2015.04.002.

Ianelli, J., Fissel, B., Stienessen, S., Honkalehto, T., Siddon, E., and Allen Akselrud, C. 2021. Assessment of the walleye pollock stock in the Eastern Bering Sea. Chapter 1. NPFMC Bering Sea and Aleutian Islands SAFE. 171 p.

IPHC. 2021. Report of the 19th Session of the IPHC Scientific Review Board (SRB019). Meeting held electronically, 21-23 September 2021. IPHC-2021-SRB019-R. 25 p.

IPHC. 2022. Report of the 98th session of the IPHC annual meeting (AM098). Electronic meeting, 24-28 January 2022. IPHC-2022-AM098-R. 60 p.

Karp, M.A., Blackhart, K., Lynch, P.D., Deroba, J., Hanselman, D., Gertseva, V., Teo, S., Townsend, H., Williams, E., and Yau, A. 2018. Proceedings of the 13th National Stock Assessment Workshop: Model Stability, Model Complexity, & Ensemble Modelling. U.S. Dept. of Commerce, NOAA. NOAA Tech. Memo. NMFS-F/SPO-189. 49 p.

Kong, T., Tran, H., and Prem, C. 2022. Fisheries data overview (2021). IPHC-2022-AM098-06 Rev_1. 15 p.

Li, Y., Bence, J.R., and Brenden, T.O. 2014. An evaluation of alternative assessment approaches for intermixing fish populations: a case study with Great Lakes lake whitefish. ICES Journal of Marine Science **72**(1): 70-81. doi:10.1093/icesjms/fsu057.

Magnusson, A., Punt, A.E., and Hilborn, R. 2012. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish and Fisheries. doi:10.1111/j.1467-2979.2012.00473.x.

Mangel, M., MacCall, A.D., Brodziak, J., Dick, E.J., Forrest, R., Pourzand, R., and Ralston, S. 2013. A perspective on steepness, reference points, and stock assessment. Canadian Journal of Fisheries and Aquatic Sciences **70**(6): 930-940.

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J.R., and Francis, R.C. 1997. A Pacific interdecadal climate oscillation with impacts on salmon production. Bulletin of the American Meteorological Society **78**(6): 1069-1079.

Martell, S., and Stewart, I. 2014. Towards defining good practices for modeling time-varying selectivity. Fisheries Research **158**: 84-95. doi:10.1016/j.fishres.2013.11.001.

Maunder, M.N., Crone, P.R., Valero, J.L., and Semmens, B.X. 2015. Growth: theory, estimation, and application in fishery stock assessment models. CAPAM Workshop Series Report 2. 57 p.

McAllister, M.K., and Ianelli, J.N. 1997. Bayesian stock assessment using catch-age data and the sampling - importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences **54**: 284-300.

McGilliard, C.R., and Palsson, W. 2021. Assessment of the rex sole stock in the Gulf of Alaska. NPFMC Gulf of Alaska SAFE. Chapter 6. 83 p.

McGilliard, C.R., Palsson, W., Havron, A., and Zador, S. 2019. Assessment of the deepwater flatfish stock complex in the Gulf of Alaska. Chapter 5. NPFMC Gulf of Alaska SAFE. 82 p.

McGilliard, C.R., Punt, A.E., Methot, R.D., Hilborn, R., and Jacobson, L. 2014. Accounting for marine reserves using spatial stock assessments. Canadian Journal of Fisheries and Aquatic Sciences: 1-19. doi:10.1139/cjfas-2013-0364.

Methot Jr, R.D., Wetzel, C.R., Taylor, I.G., and Doering, K. 2020a. Stock synthesis user manual version 3.30.16. NOAA Fisheries. 225 p.

Methot Jr, R.D., Wetzel, C.R., Taylor, I.G., and Doering, K. 2020b. Stock Synthesis User Manual Version 3.30.15. U.S Department of Commerce, NOAA Processed Report NMFS-NWFSC-PR-2020-05. 226 p.

Methot Jr, R.D., Wetzel, C.R., Taylor, I.G., Doering, K.L., and Johnson, K.F. 2021a. Stock synthesis user manual version 3.30.18. NOAA Fisheries. Seattle, WA. 234 p.

Methot Jr, R.D., Wetzel, C.R., Taylor, I.G., Doering, K.L., and Johnson, K.F. 2021b. Stock synthesis user manual version 3.30.17. NOAA Fisheries, Seattle, WA. 238 p.

Methot, R.D., and Taylor, I.G. 2011. Adjusting for bias due to variability in estimated recruitments in fishery assessment models. Canadian Journal of Fisheries and Aquatic Sciences **68**: 1744-1760.

Methot, R.D., and Wetzel, C.R. 2013a. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research **142**(0): 86-99. doi:http://dx.doi.org/10.1016/j.fishres.2012.10.012.

Methot, R.D., and Wetzel, C.R. 2013b. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Appendix A: Technical description of the Stock Synthesis assessment program. Fisheries Research **142**: 26 p.

Methot, R.D., Wetzel, C.R., and Taylor, I.G. 2019. Stock Synthesis User Manual Version 3.30.13. NOAA Fisheries. Seattle, WA. 213 p.

Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. ICES Journal of Marine Science: Journal du Conseil **56**(4): 473-488. doi:10.1006/jmsc.1999.0481.

Monnahan, C.C., and Stewart, I.J. 2017. The effect of hook spacing on longline catch rates: Implications for catch rate standardization. Fish. Res. doi:10.1016/j.fishres.2017.10.004.

Monnahan, C.C., and Haehn, R. 2020. Assessment of the flathead sole-Bering flounder complex in the Bering Sea and Aleutian Islands. Chapter 9. NPFMC Bering Sea and Aleutian Islands SAFE. 91 p.

Ormseth, O.A. 2021. Assessment of the Alaska plaice stock in the Bering Sea and Aleutian Islands. Chapter 10. NPFMC Bering Sea and Aleutian Islands SAFE. 63 p.

Pennington, M., and Volstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics **50**: 725-732.

Peterson, C.D., Gartland, J., and Latour, R.J. 2017. Novel use of hook timers to quantify changing catchability over soak time in longline surveys. Fisheries Research **194**: 99-111. doi:10.1016/j.fishres.2017.05.010.

Peterson, M.J., and Carothers, C. 2013. Whale interactions with Alaskan sablefish and Pacific halibut fisheries: Surveying fishermen perception, changing fishing practices and mitigation. Marine Policy **42**: 315-324. doi:10.1016/j.marpol.2013.04.001.

Peterson, M.J., Mueter, F., Criddle, K., and Haynie, A.C. 2014. Killer whale depredation and associated costs to Alaskan sablefish, Pacific halibut and Greenland turbot longliners. PloS one **9**(2): e88906. doi:10.1371/journal.pone.0088906.

Piner, K.R., and Wischnioski, S.G. 2004. Pacific halibut chronology of bomb radiocarbon in otoliths from 1944 to 1981 and a validation of ageing methods. Journal of Fish Biology **64**: 1060-1071.

Punt, A.E., Haddon, M., and Tuck, G.N. 2015. Which assessment configurations perform best in the face of spatial heterogeneity in fishing mortality, growth and recruitment? A case study based on pink ling in Australia. Fisheries Research **168**: 85-99. doi:10.1016/j.fishres.2015.04.002.

Ralston, S. 2005. An assessment of starry flounder off California, Oregon, and Washington. Pacific Fishery Management Council. 86 p.

Sampson, D.B., Hamel, O.S., Bosley, K., Budrick, J., Cronin-Fine, L., Hillier, L.K., Hinton, K.E., Krigbaum, M.J., Miller, S., Privitera-Johnson, K.M., Ramey, K., Rodomsky, B.T., Solinger, L.K., and Whitman, A.D. 2017. 2017 assessment update for the US west coast stock of arrowtooth flounder. Pacific Fisheries Management Council, Portland, OR. 152 p.

Shotwell, K., Spies, I., Ianelli, J.N., Aydin, K., Hanselman, D.H., Palsson, W., Siwicke, K., Sullivan, J., and Yasumiishi, E. 2021. Assessment of the arrowtooth flounder stock in the Gulf of Alaska. Chapter 7. NPFMC Gulf of Alaska SAFE. 99 p. .

Shotwell, S.K., Spies, I., Britt, L., Bryan, M., Hanselman, D.H., Nichol, D.G., Hoff, J., Palsson, W., Wilderbuer, T.K., and Zador, S. 2020. Assessment of the arrowtooth flounder stock in the Eastern Bering Sea and Aleutian Islands. Chapter 6. NPFMC Bering Sea and Aleutian Islands SAFE. 88 p.

Spies, I., Haehn, R., Siddon, E., Conner, J., Markowitz, E., Yeung, C., and Ianelli, J. 2021. Assessment of the yellowfin sole stock in the Bering Sea and Aleutian Islands. NPFMC Bering Sea and Aleutian Islands SAFE. 103 p.

Starr, P.J. 2009a. Petrale sole (*Eopsetta jordani*) in British Columbia: stock assessment for 2006/07 and advice to managers for 2007/08. Can. Sci. Advis. Sec. Res. Doc. 2009/070. 140 p.

Starr, P.J. 2009b. English sole (*Parophrys vetulus*) in British Columbia, Canada: stock assessment for 2006/07 and advice to managers for 2007/08. Can Sci. Advis. Sec. Res. Doc. 2009/069. 155 p.

Stewart, I., and Hicks, A. 2019a. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2018. IPHC-2019-AM095-09. 26 p.

Stewart, I., and Webster, R. 2019. Overview of data sources for the Pacific halibut stock assessment, harvest policy, and related analyses. IPHC-2019-AM095-08. 76 p.

Stewart, I., and Hicks, A. 2019b. 2019 Pacific halibut (*Hippoglossus stenolepis*) stock assessment: development. IPHC-2019-SRB014-07. 100 p.

Stewart, I., and Hicks, A. 2020. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2019. IPHC-2020-SA-01. 32 p.

Stewart, I., and Hicks, A. 2021. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2020. IPHC-2021-SA-01. 33 p.

Stewart, I., and Webster, R. 2022. Overview of data sources for the Pacific halibut stock assessment, harvest policy, and related analyses. IPHC-2022-SA-02. 56 p.

Stewart, I., and Hicks, A. 2022. Assessment of the Pacific halibut (*Hippoglossus stenolepis*) stock at the end of 2021. IPHC-2022-SA-01. 36 p.

Stewart, I., Hicks, A., and Hutniczak, B. 2021. Evaluation of directed commercial fishery size limits in 2020. IPHC-2021-AM097-09. 28 p.

Stewart, I.J. 2007. Updated U.S. English sole stock assessment: mstatus of the resource in 2007. Pacific Fishery Management Council. 213 p.

Stewart, I.J. 2014. Overview of data sources for the Pacific halibut stock assessment and related analyses. IPHC Report of Assessment and Research Activities 2013. p. 95-168.

Stewart, I.J., and Martell, S.J.D. 2014. A historical review of selectivity approaches and retrospective patterns in the Pacific halibut stock assessment. Fisheries Research **158**: 40-49. doi:10.1016/j.fishres.2013.09.012.

Stewart, I.J., and Hamel, O.S. 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Canadian Journal of Fisheries and Aquatic Sciences **71**(4): 581-588. doi:10.1139/cjfas-2013-0289.

Stewart, I.J., and Martell, S.J.D. 2015. Reconciling stock assessment paradigms to better inform fisheries management. ICES Journal of Marine Science **72**(8): 2187-2196. doi:10.1093/icesjms/fsv061.

Stewart, I.J., and Martell, S.J.D. 2016. Appendix: Development of the 2015 stock assessment. IPHC Report of Assessment and Research Activities 2015. p. A1-A146.

Stewart, I.J., and Monnahan, C.C. 2017. Implications of process error in selectivity for approaches to weighting compositional data in fisheries stock assessments. Fisheries Research **192**: 126-134. doi:10.1016/j.fishres.2016.06.018.

Stewart, I.J., and Hicks, A.C. 2018. Interannual stability from ensemble modelling. Canadian Journal of Fisheries and Aquatic Sciences **75**: 2109-2113. doi:10.1139/cjfas-2018-0238.

Stewart, I.J., Wallace, J.R., and McGilliard, C.R. 2009. Status of the U.S. yelloweye rockfish resource in 2009. Pacific Fishery Management Council, Portland, OR. 236 p.

Stewart, I.J., Leaman, B.M., Martell, S., and Webster, R.A. 2013a. Assessment of the Pacific halibut stock at the end of 2012. IPHC Report of Assessment and Research Activities 2012. p. 93-186.

Stewart, I.J., Hicks, A.C., Taylor, I.G., Thorson, J.T., Wetzel, C., and Kupschus, S. 2013b. A comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. Fisheries Research **142**(0): 37-46. doi:http://dx.doi.org/10.1016/j.fishres.2012.07.003.

Stokes, K. 2019. Independent peer review for the 2019 IPHC stock assessment. August 2019. 31 p. https://www.iphc.int/uploads/pdf/sa/2019/stokes_2019-independent_peer_review_for_the_2019_iphc_stock_assessment.pdf.

Taylor, I.G., Doering, K.L., Johnson, K.F., Wetzel, C.R., and Stewart, I.J. 2021. Beyond visualizing catch-at-age models: Lessons learned from the r4ss package about software to support stock assessments. Fisheries Research **239**. doi:10.1016/j.fishres.2021.105924.

Thompson, G.G., and Lauth, R.R. 2012. Chapter 2: Assessment of the Pacific cod stock in the Eastern Bering Sea and Aleutian Islands area. *In* NPFMC Bering Sea and Aleutian Islands SAFE. p. 245-544.

Thorson, J.T. 2019. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fisheries Research ***217***: 133-139.

Thorson, J.T., Johnson, K.F., Methot, R.D., and Taylor, I.G. 2016. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. doi:10.1016/j.fishres.2016.06.005.

Turnock, B.J., McGilliard, C.R., and Palsson, W. 2017. Assessment of the flathead sole stock in the Gulf o0f Alaska. Chapter 8. NPFMC Gulf of Alaska SAFE. 72 p.

Ualesi, K., Wilson, D., Jones, C., Rillera, R., and Jack, T. 2022. IPHC Fishery Independent Setline Survey (FISS) design and implementation in 2021. IPHC-2022-AM098-07. 13 p.

Warton, D.I., Thibaut, L., and Wang, Y.A. 2017. The PIT-trap-A "model-free" bootstrap procedure for inference about regression models with discrete, multivariate responses. PloS one **12**(7): e0181790. doi:10.1371/journal.pone.0181790.

Waterhouse, L., Sampson, D.B., Maunder, M., and Semmens, B.X. 2014. Using areas-as-fleets selectivity to model spatial fishing: Asymptotic curves are unlikely under equilibrium conditions. Fisheries Research **158**: 15-25. doi:10.1016/j.fishres.2014.01.009.

Webster, R. 2018. Space-time modelling of fishery independent setline survey data. IPHC-2018-AM094-07. 27 p.

Webster, R. 2021. IPHC Fishery-Independent Setline Survey (FISS) and commercial data modelling. IPHC-2021-SRB019-05. 43 p.

Webster, R. 2022. Space-time modelling of survey data. IPHC-2022-AM098-08. 6 p.

Webster, R., and Stewart, I. 2022. Revision of the IPHC length-weight relationship. IPHC-2022-AM098-INF07. 29 p.

Webster, R.A., Clark, W.G., Leaman, B.M., and Forsberg, J.E. 2013. Pacific halibut on the move: a renewed understanding of adult migration from a coastwide tagging study. Canadian Journal of Fisheries and Aquatic Sciences **70**(4): 642-653. doi:10.1139/cjfas-2012-0371.

Webster, R.A., Soderlund, E., Dykstra, C.L., and Stewart, I.J. 2020. Monitoring change in a dynamic environment: spatio-temporal modelling of calibrated data from different types of fisheries surveys of Pacific halibut. Canadian Journal of Fisheries and Aquatic Sciences **77**: 1421-1432.

Wetzel, C.R. 2019. Status of petrale sole (*Eopsetta jordani*) along the U.S. west coast in 2019. Pacific Fishery Management Council, 7700 Ambassador Place NE, Suite 101, Portland, OR 97220. 256 p.

Wetzel, C.R., and Berger, A.M. 2021. Status of Dover sole (*Microstomus pacificus*) along the U.S. West Coast in 2021. Pacific Fisheries Management Council, Portland, Oregon. 289 p.

Appendices

Appendix A: Supplementary material

In addition to this document, supplementary material is available electronically, including:

1) Stock synthesis input files for each of the assessment models included in the proposed ensemble: data file, weight-at-age file, control file with model configuration, starter and forecast files with additional settings. Each of these files has been extensively annotated to aid in locating the various sections, as well as identifying which options and features were implemented or are irrelevant for the configuration.

2) Output from each of the stock assessment models: a sub-directory of all plotting and diagnostic output from each model created by the r4ss package (the entire set can be loaded at once via opening the "_SS_output.html" file), and the raw report (text) file from each model. The report file has not been annotated and contains some information not relevant to the Pacific halibut model configurations; content and formats can be determined from the stock synthesis user manual (Methot Jr et al. 2021b) and technical documentation (Methot and Wetzel 2013b).

3) Copies of the primary software documentation including the general modelling approach implemented in stock synthesis (Methot and Wetzel 2013a), the technical documentation (Methot and Wetzel 2013b) and the current user manual (Methot Jr et al. 2021b). From these documents, detailed model equations, data configurations, and control settings can be evaluated for the specific features implemented in the models for Pacific halibut.

4) The overview of data sources (Stewart and Webster 2022) and the stock assessment results (Stewart and Hicks 2022) from the 2021 stock assessment.

5) The documentation from the development of the most recent (2019) full stock assessment (Stewart and Hicks 2019b).

6) Recent relevant IPHC manuscripts describing the bootstrapping method employed for fishery and FISS age compositions (Stewart and Hamel 2014), the history of the halibut stock assessment (Stewart and Martell 2014), an evaluation of data weighting and process-error considerations (Stewart and Monnahan 2017), the general rationale for the ensemble approach (Stewart and Martell 2015), and the stability properties of ensemble assessments (Stewart and Hicks 2018).

7) A full record of the historical stock assessment documentation from 1978 to the present can be found on the IPHC's web site (https://www.iphc.int/management/science-and-research/stock-assessment). Individual Scientific Review Board reports and presentations (2013-2022) are available through the IPHC's meetings webpage (https://www.iphc.int/iphc-meetings).