B. GEORGES BANK HADDOCK

List of Attendees of Working Group Webinars

NAME	AFFILIATION
Brian Linton*	NMFS
Charles Perretti*	NMFS
Jamie Cournane*	NEFMC
Kathryn Cooper-MacDonald*	DFO
Kevin Friedland*	NMFS
Liz Brooks*	NMFS
Monica Finley*	DFO
Scott Large*	NMFS
Steve Cadrin*	SMAST
Tim Barrett*	DFO
Tom Carruthers*	Blue Matter Science
Yanjun Wang*	DFO
Abigail Tyrell	NMFS
Alex Dalton	DFO
Alex Hansell	NMFS
Alexander Dunn	NMFS
Andrew Jones	NMFS
Barry Gibson	Stakeholder
Brian Stock	NMFS
Briony Donahue	Maine DMR
Brooke Lowman	NMFS
Catriona Regnier-McKellar	DFO
Chelsey Karbowski	Stakeholder
Chris Legault	NMFS
Cole Carrano	SMAST
Cyril Boudreau	Nova Scotia DFA
Daniel Salerno	Stakeholder
Dave McElroy	NMFS
Devin Archibald	-
Ed Cunnie	Stakeholder
Gene Bergson	Stakeholder
Hank Soule	Stakeholder
Jessica Blaylock	NMFS
Jim Manning	NMFS
Keith Hankowsky	SMAST
Kelly Kraska	DFO

Kevin Niland	Stakeholder
Kris Vascotto	Stakeholder
Kyle Molton	NMFS
Larry Alade	NMFS
Libby Etrie	NEFMC
Lottie Bennett	DFO
Marina Cucuzza	NMFS
Mark Terceiro	NMFS
Mark Wuenschel	NFMS
Max Grezlik	SMAST
Melissa Errend	NEFMC
Michael Pierdinock	Stakeholder
Michael Plaia	Stakeholder
Michele Traver	NMFS
Michelle Greenlaw	DFO
Mike Simpkins	NMFS
Paul Nitschke	NMFS
Quanh Huynh	Blue Matter Science
Quinn McCurdy	DFO
Rebecca Peters	Maine DMR
Ricky Tabandera	NMFS
Robin Frede	NEFMC
Russell Brown	NMFS
Ryan Morse	NMFS
Spencer Talmage	NMFS
Steven Devitt	Stakeholder
Susan Wigley	NMFS
Tara McIntyre	DFO
Tara Trinko-Lake	NMFS
Tim Miller	NMFS
Tom Nies	NEFMC
Vitalii Sherement	WHOI

*Working group members

Executive Summary

TOR 1: Review existing research efforts, data, and habitat information in the Gulf of Maine and Georges Bank, identify any findings relevant to influences of ecosystem conditions on haddock, and consider those findings, as appropriate, in addressing other TORs. For processes that the working group deems important and promising that are not currently feasible to consider quantitatively, describe next steps for development, testing, and review of quantitative relationships and how they could best inform assessments.

The working group (WG) developed habitat distribution models to examine the extent to which ecosystem variables could explain changes in haddock abundance and distribution for both Gulf of Maine (GOM) and Georges Bank (GB) haddock. Habitat models were developed using machine learning methodology and applied to data from 1976-2019. Haddock habitat scores have increased over the past two decades, with the most dramatic increases in the past six years, coinciding with the strong 2013 year class. The model indicates that, in recent years, habitat appears rather homogenous across the GOM/GB stock areas. The WG discussed the possibility that the habitat scores may be driven by changes in haddock distribution, as opposed to habitat driving haddock distribution. Therefore, it was unclear whether the model had established a causal relationship between the selected habitat variables and haddock distribution.

Potential next steps discussed by the WG include developing a mechanistic understanding of how the model-selected habitat variables drive changes in haddock distribution, and verifying habitat model results of changes in seasonal movement in recent years.

The working group reviewed another analysis which aimed to model spatial distributions of haddock on GB using a Poisson-link delta model composed of encounter-probability and catch rates with a process structure of spatial, temporal and spatiotemporal components. Model predicted Area of Occupancy of haddock on GB varied substantially over time, with consistent trends between spring and fall. Density-dependence made the greatest contribution to the variations in both seasons.

TOR 2: Estimate catch from all sources including landings and discards. Describe the spatial and temporal distribution of landings, discards, and fishing effort. Characterize the uncertainty in these sources of data.

US commercial landings data are summarized for the year range where digital records are available (1964-present). Annual landings peaked at over 50,000 mt in the 1960s, dropped precipitously in the 1970s, followed by a brief peak of 15,000-20,000 mt per year in the early 1980s. Landings reached an all-time low in the 1990s, with the stock being declared collapsed in 1992. Rebuilding occurred in the late 1990s, and since then annual landings have been around 2,000-6,000 mt. The primary gear has been otter trawlers, with very minor contributions from gillnet and hook/line gears.

US commercial discards were re-estimated for the years that observer data are available (1989-2019). Discards peaked in 1994, a period of strict landings regulations aimed at rebuilding haddock; discards were estimated at over 2,000 mt (43% of total catch that year). Discards peaked again in 2006 and 2007, in large part due to slowed growth from the 2003 year class leading to many undersized fish. An emergency rule was enacted to reduce the minimum size. Discards peaked again in 2013-2016 as the 2010 and 2013 year classes reached the fishery. The precision of discards was low at the start of the observer program (CV ranged from 0.27-1.6 from 1989-1998), but has improved substantially, with a CV of 20% or less in most years since then.

Canadian commercial landings were re-estimated for years 1987-2019 due to concerns about the length-weight equation. As with US landings, Canadian landings on eastern Georges Bank track large year classes. The fishery lands fish at ages 1 and 2, whereas these ages are discarded in the US fleet. Since the mid-1990s, there are proportionally fewer fish landed below age 3.

Canadian commercial discard estimates from the Canadian scallop fleet are greatest at age 0, and persist though about age 4 or 5 in small amounts. Discards are greatest on the strong year classes, which track well through the 2003 year class at age 1 (in 2004). The magnitude of discards is remarkably small in years since 2005, particularly given the exceptional size of the 2010 and 2013 year classes.

TOR 3: Present the survey data being used in the assessment (e.g., indices of relative or absolute abundance, recruitment, state surveys, age-length data, etc.). Characterize the uncertainty in these sources of data.

The NEFSC bottom trawl survey (BTS) transitioned to a new vessel, the R/V Henry B. Bigelow, in 2009. Paired tows with the Albatross IV took place in 2008 during the spring and fall BTS and also included site-specific tows in the summer (636 total paired-gear sets; NEFSC Vessel Calibration Working Group 2007; Miller et al. 2010). New vessel calibration factors were used for this assessment. The new calibration factors were developed by Miller (2013), which examined 16 species in a hierarchical mixed effect framework that allowed for variation among and between paired tows. The impact of the Miller (2013) length-based calibration compared to the current Brooks et al. (2010) calibration is most discernible at the smallest sizes and in tracking the exceptional year classes.

The NEFSC BTS indices of abundance and biomass were high at the beginning of the fall (1963) survey time series as a result of an exceptional 1963 yearclass. By the time the spring survey began (1968), most of that cohort had been fished out of the population. Over the next 30 years, population increases were driven by stronger than average year classes in 1975, 1978, and 1980, 1985, 1998, and 2000. In the last 20 years, biomass reached the highest levels ever observed, which was driven by 3 boomer year classes: 2003, 2010, and 2013.

The Canadian Department of Fisheries and Oceans (DFO) has conducted an annual bottom trawl survey on Georges Bank (statistical areas 5Zj, 5Zm, 5Zh, 5Zn, 5Zg, and 5Zo) between the 6th week of the year (Mid-February) until the 13th week (end of March), since 1987. The DFO survey biomass index was low in 1987 and remained low through much of the 1990's until the appearance of the strong 1998 year class. As seen in the NEFSC surveys, the biomass reached the highest levels observed during the last 20 years, which was driven by 3 boomer year classes: 2003, 2010, and 2013.

TOR 4: Estimate annual fishing mortality, recruitment and stock biomass (both total and spawning stock) for the time series, and estimate their uncertainty. Compare the time series of these estimates with those from the previously accepted assessment model, and evaluate the strength and direction of any retrospective pattern(s) in both the current and the previously accepted model. Enumerate possible sources of the retrospective patterns and characterize plausibility, if possible.

This TOR was addressed by updating the current model (Virtual Population Assessment, VPA) with new data to the extent possible, examining the impact of new data treatments on the scale of abundance and the measure of retrospective pattern (Mohn's rho). Next, a bridge was built between the VPA and a statistical catch at age model (Age Structured Assessment Program, ASAP; Legault and Restrepo, 1998), although model structure, diagnostics, fleet selectivity blocking, and post-hoc weighting adjustments were examined for ASAP runs. This bridge was primarily a stepping stone en route to modeling the proposed base model in a state space framework (Woods Hole Assessment Model, WHAM; Stock and Miller, 2020). A full exploration of model configuration and model building was pursued in arriving at the proposed base model in WHAM.

The base model exhibits a retrospective pattern, but rho-adjusted estimates of SSB and average F in 2019 are within the 95% confidence interval. There are various potential causes for the retrospective pattern, including changes in selectivity due to changes in growth, smearing of age classes around the recent exceptionally strong year classes, and possible changes in natural mortality over time.

TOR 5: Update or redefine status determination criteria (SDC point estimates or proxies for BMSY, BTHRESHOLD, FMSY and MSY) and provide estimates of their uncertainty. If analytic model-based estimates are unavailable, consider recommending alternative measurable proxies for BRPs.

The proposed base model is in WHAM, which has built-in projection capabilities. In calculating the reference points, WHAM can propagate parameter uncertainty into the F_MSY and B_MSY proxies. U.S. national standard guidelines specify that MSY reference points or proxies should reflect "prevailing ecological, environmental conditions and fishery technological characteristics (e.g., gear selectivity), and the distribution of catch among fleets" (NOAA 2016). The working group agreed to use the full time series of recruitment estimates and a recent 5 year average of selectivity, weights, and maturity at age. Use of the full time series of recruitment is justified by examining the trend in average recruitment over different time intervals, from a recent 5 year average to the full time series, in increments of 5 years. Short term average recruitment is strongly influenced by the recent exceptional year classes, while windows from the most recent 35 years through the full time series show a fairly stable mean recruitment and very little difference in the empirical cumulative distribution (cdf). For selectivity, examining the mean age-specific value over the same temporal windows as average recruitment, there is very little difference in selectivity at ages 1 and 6-9+. Selectivity for ages 2-5 shows increasing values from the longer the time series. The shorter window seems more appropriate and reflects the impact of both changes in growth and changes in management (e.g., minimum size). For these reasons, a 5 year recent average appears justifiable. For similar reasons, a recent 5 year average is more reflective of population growth and would likely make a more reasonable basis for calculating YPR, SSB/R. The maturity ogive has 2 stanzas, and the final stanza is 4 years; for consistency, a 5 year average is recommended for maturity at age.

TOR 6: Define the methodology for performing short-term projections of catch and biomass under alternative harvest scenarios, including the assumptions of fishery selectivity, weights at age, maturity, and recruitment.

Methodology for short term projections for Georges Bank haddock has been adjusted in nearly every management update since 2012 in order to deal with changes in growth that impact projected weights at age and selectivity at age. For this research track assessment benchmark, assumptions for weights at age were re-visited. Analyses conducted for this assessment suggest that using a 2 year average of recent weights at age (up to age 8 at least) is the most robust assumption for short-term projections. For the plus group, the recommendation for projection weights at age 9 for exceptional year classes is to use the minimum of a 2 year average or the mean ratio of age 9 to age 8 (1.06 or 1.05); for other year classes, a 2 year average for plus group weight is expected to perform as well as any other approach.

Analyses about trends in selectivity by the window of years averaged over (see TOR5), suggest that the current 5 year average will capture relevant recent exploitation characteristics. As the selectivity is logistic, no special assumption is needed for the oldest ages.

Numbers at age in the base model have a 2DAR1 autoregressive structure on random effects, and the working group agreed to let the model propagate these into the future. The NAA deviations at the end of the model are not large, and revert to zero within 2-3 years.

TOR 7: Review, evaluate and report on the status of the Stock Assessment Review Committee (SARC) and Working Group research recommendations listed in most recent SARC reviewed assessment and review panel reports. Identify new research recommendations.

The WG reviewed progress towards meeting the research recommendations from the GARM III GB haddock assessment (NEFSC 2008), which was the last benchmark assessment for the GB haddock stock, and from the 2019 GOM haddock management track assessment (NEFSC in preparation), which was the most recent assessment of the GB haddock stock. The WG proposed new research recommendations, including updating the gutted:whole weight conversion factors, exploring methods for estimating weight at age and fishery selectivity for strong year classes in short term projections, and incorporating density-dependent processes (e.g., growth, spatial distribution, and natural mortality) into stock assessment models.

TOR 8: Develop a "Plan B" for use if the accepted assessment model fails in the future.

In the event that the proposed base model (WHAM_BASE) or alternative model configurations are not accepted as a basis for status determination and fishery management advice, the proposed alternative assessment framework is to use the Plan B Smooth. This has been the proposed back-up method since the 2017 management track assessment (NEFSC 2017), but has never needed to be applied. A research track working group investigating performance of index-based methods (NEFSC 2020 in prep; Legault et al., submitted), conducted a large-scale simulation study to understand if simpler methods would perform better in situations where an age-based assessment was rejected due to a large retrospective pattern (Mohn's rho of about 0.5 for SSB). Plan B smooth was one of the methods, and across all scenarios considered, its performance was generally robust, except for scenarios where retrospective patterns were caused by unreported catch and the stock was depleted. The Georges Bank haddock stock is not depleted.

Plan B smooth fits a loess (second degree polynomial) to average survey biomass, and then a log-linear regression is fit to the last 3 years of the loess fit. The slope from the log-linear regression is then exponentiated and becomes a multiplier on either catch or quota.

TOR 9: Review and present any research related to recruitment processes (e.g., spawning and larval transport, and retention), and potential hypotheses for large recruitment events.

The Working Group considered an update to the analysis of the effect of the fall bloom on haddock recruitment. It has been hypothesized that the recruitment of haddock on Georges Bank (GB) may be influenced by the provisioning effects of the fall bloom on pre-spawning adults (Friedland et al. 2015). The updated analysis tested the patency of the relationship between GB haddock recruitment and the dimensions of phytoplankton blooms forming on GB in the fall prior to spawning (Friedland 2021) . The analysis was successful in modeling the exceptional recruitment in 2013 and can thus be attributed to the provisioning hypothesis and more generally showed that relationship holds with a more than a doubling of recruitment data considered in the original analysis. The Working Group concluded that the provisioning hypothesis may be contributing to the observed pattern of recruitment of GB haddock. Although there were three exceptionally strong year classes in the last two decades, only two of them (2003 and 2013) were identified to have high bloom magnitude. The bloom algorithm did not find a clean beginning and end to a bloom event associated with the 2010 year class.

The WG also considered ongoing research by Sheremet et al. (In prep.) on the retention of haddock eggs and larvae based on both real and simulated ocean drifter trajectories. Sheremet et al. (In Prep.) used the Finite Volume Coastal Ocean Model (FVCOM; Chen et al. 2006), a prognostic, unstructured-grid ocean circulation model that assimilates hydrography, altimetry, atmospheric conditions, and tides, with hourly output, to study patterns and variability in retention. For a weekly time step, 10,000 theoretical drifters, which serve as a proxy for spawned eggs, were released from the Northeast peak of Georges Bank and their trajectory was projected based on depth-averaged flows. For drifters released on the NE peak of Georges Bank, the yearly climatology showed no retention while the monthly climatology showed that spawning in April would produce the highest retention. It was concluded that high retention in general does not guarantee strong recruitment; spawning biomass is necessary to produce eggs, and after the egg/larval stage, many other factors can act post settlement. However, it is certainly true that high recruitment cannot occur with low retention.

These same analyses conducted for the Northeast Peak were repeated for releases in the Great South Channel, known to be another spawning location for haddock. As with releases on the Northeast Peak, retention was greatest for release in April.

Lastly, an examination of real drifter tracks produced results that align with the simulated drifter results, and other information reviewed in TOR12.

TOR 10: Review and present any research related to density-dependent growth.

Density-dependent growth of GB haddock was previously reported in several early research studies (Clark et al. 1982; Ross and Nelson 1992; Brodziak and Link 2008; Brodziak et al. 2008). A dramatic increase in haddock biomass from a few episodic recruitments with a precipitous decrease in somatic growth in the most recent decade has caused fishery managers and stakeholders great concern, particularly for haddock concentrating on EGB (TRAC 2017). The WG reviewed an analysis primarily focused on EGB. A 2-step Generalized Additive Model (GAM) regression approach was applied to investigate the mechanistic response of somatic growth in haddock to environmental and intraspecific competition (see details in Wang et al., 2021). Density-dependence including both cohort and year effect was identified as the most important factors in explaining haddock growth changes.

TOR 12: Review data related to stock structure of haddock on Georges Bank (including Eastern Georges Bank management area) and implications for assessments conducted on the whole bank and on subareas of the bank.

Stock identity of haddock on Georges Bank remains uncertain, because stock boundaries are indistinct and geographic stock structure appears to be largely influenced by haddock abundance. Large channels form partial barriers to movement of juvenile and adult haddock, and larval retention gyres also limit mixing of eggs and larvae across these channels. During periods of low abundance, the spatial distribution of juvenile and adult haddock was more discontinuous, with discrete concentrations on eastern Georges Bank, in the Great South Channel, in the Gulf of Maine and on the Scotian Shelf. However, during periods of high abundance, spatial distributions were more continuous between eastern Georges Bank and the Great South Channel. Therefore, static boundaries that meet the unit stock assumption are difficult to define.

Geographic variation in genetic, phenotypic and demographic traits suggests an isolation-by-distance pattern, with occasional mixing of haddock from multiple areas. The US Georges Bank Management unit includes multiple discrete spawning components of haddock on eastern Georges Bank and the adjacent Great South Channel. During a period of low abundance in the late 1980s-early 1990s, juvenile and adult haddock were discretely distributed on eastern Georges Bank and had faster growth and maturity rates than haddock in the Great South Channel.

In summary, there are no persistent stock boundaries of haddock on Georges Bank because the resource distribution and connectivity among areas are dynamic. Therefore, previous boundaries defined during periods of low abundance may need to be reconsidered. Length and age compositions of haddock were similar among eastern Georges Bank, western Georges Bank, and the Great South Channel, suggesting a well-mixed stock. However, exploratory stock assessment modeling did not indicate evidence of emigration from eastern Georges Bank. Separate assessment of eastern Georges Bank haddock includes a relatively homogeneous resource but does not account for larger-scale recruitment dynamics, and emigration of juveniles and adults to western Georges Bank may be a factor in recent retrospective patterns. The current distribution and connectivity of haddock across the Bank suggest that haddock on Georges Bank is a single stock.

Introduction

Assessment History

The Georges Bank haddock (*Melanogrammus aeglefinus*) stock assessment has involved age-based analyses since the mid 1900s (e.g., Schuck 1949) and Virtual Population Analysis (VPA) since the 1960s (e.g., Hennemuth 1969). Calibrated VPAs for Georges Bank haddock (e.g., Clark et al. 1982) and eastern Georges Bank haddock (e.g., Gavaris 1989) were updated regularly through the Northeast Stock Assessment Workshop (e.g., NEFSC 1992) and Transboundary Resources Assessment Committee, TRAC (e.g., DFO 1998). The 1992 and 1995 assessments of Georges Bank haddock concluded that the stock was collapsed (NEFSC 1992, 1995), and subsequent assessments monitored stock rebuilding (NEFSC 1997, 1998, 2002). A benchmark assessment was performed in 2007-2008, and since then 4 update assessments have been performed, most recently in 2019, at which time it was determined that the stock was not overfished and overfishing was not occurring (NEFSC 2019). The eastern Georges Bank haddock assessment was updated in 2019 but was not acceptable because of a large retrospective pattern (Mohn's rho of 1.75 on SSB and -0.677 on average F; TRAC 2019), so subsequent assessments have been empirical (TRAC 2021).

The most recent benchmark assessment of Georges Bank haddock occurred in 2008 at the Groundfish Assessment Review Meeting (GARM-III). For this meeting, both a VPA and a statistical catch at age model (Age Structured Assessment Protocol, ASAP; Legault and Restrepo 1998) were put forward as assessment models. The review panel deemed that the VPA was the preferred model, with acceptable diagnostics and no retrospective pattern of concern, while the ASAP model was considered preliminary, with specific mention of further exploration of number and placement of fleet selectivity blocks to account for changes in growth and changes in the fishery. It was recommended that the full time series of catch back to 1931 should be included. The value of rho was 0.07 for SSB and -0.07 for average F. No retrospective adjustment was made. The spawning stock biomass in 2007 was estimated to be 315,976 mt (CV=0.2) and average F on ages 5-7 was 0.23 (cv=0.16). The stock had been rebuilding from an overfished status and with these results was declared rebuilt.

Management update assessments occurred in 2012, 2015, 2017, and 2019, still in the VPA framework. Salient points for these updates are:

- 2012: This assessment used data through 2010. Stock status was not overfished and overfishing was not occurring (SSB>>SSB_MSY). The initial estimate of the 2010 year class was extremely large and highly uncertain due to it only being observed in the DFO and NEFSC spring survey at age 1 in 2011 (terminal year+1 index observations). Mohn's rho values were 0.2 for SSB and -0.15 for average F, and no retrospective pattern adjustment was made. For short-term projections, the initial estimate of age 1 in 2011 (the 2010 year class) was scaled downward by 50% based on examination of the direction and magnitude of change between initial estimates of two recent above average year classes (2003 and 1998) (Figure 3). As noted at the time, "the relative change in estimates of year class size was compared between an estimate based on one year of data and model runs that added one additional year of data. Rather than computing this relative change from the terminal point (as in Mohn's rho), the relative change was calculated from the initial model estimate. This more closely replicates the present scenario: survey observations from one year are all that is being used to estimate the 2010 year class. In future years, the estimate will become less certain as data are added to the model."

- 2015: This assessment used data through 2014. Stock status was not overfished and overfishing was not occurring (SSB>>SSB_MSY). The 2013 year class was estimated to be the largest ever in the history of this stock, though it was highly uncertain due to the limited observations from which it was estimated. Mohn's rho values were 0.5 for SSB and -0.34 for average F. A rho adjustment was made to terminal year SSB and F estimates for the purpose of status determination, and to starting numbers at age (all were scaled by 0.67=1/(1+rho_SSB)) for short-term projections. Earlier years in the time series of VPA estimates are not rho adjusted, because the calculation of Mohn's rho is relative to the terminal year estimate. The decision to adjust starting numbers was based on a diagnostic plot of whether the rho-adjusted SSB and average F were outside of the 80% confidence interval of their estimates–the adjusted values were outside, which warranted the adjustment (Figure 4). The downscaling of initial year class estimates that was explored in the 2012 update were evaluated again, and with the 4 additional years of data, the downscaling of initial 2003 and 2010 year class estimates was 0.28 and 0.63. In response to this, two short-term projections were made, one that multiplied all starting ages by 0.67 (using rho_SSB), and one that multiplied ages 2-9+ by 0.67 and adjusted age 1 by 0.33 to reduce the influence of the extremely uncertain 2013 year class.

- 2017: This assessment used data through 2016. Stock status was not overfished and overfishing was not occurring (SSB>>SSB_MSY). However, calculation of average F for this assessment departed from previous assessments, in that for the first time a numbers-weighted average F was calculated to deal with what appeared to be anomalously high F on age 7 (average F is calculated over ages 5-7). In 2016, the very small 2009 year class was 7 years old, and was adjacent to the very large 2010 year class at age 6. The average F at age in 2016 for ages 5, 6, and 7 was 0.09, 0.112, and 0.516, respectively. It seemed highly likely that expanding the landings length samples by the 2016 age-length key, assigned more fish to age 7 than would be expected given the size of that year class ("age-smearing"). The mean size at ages 5-7 in the landings was 47.0, 47.7, and 50.4, respectively, with a lot of overlap in their distributions. The terminal year estimate of average F was 0.24, and after rho-adjustment was 0.656, implying overfishing on a stock for which <10% of the quota had been taken in that year. In this circumstance, a pragmatic decision was made to calculate a numbers-weighted average F to reduce the influence of the estimated F on the 2009 year class. Mohn's rho values were 0.89 for SSB and -0.55 for average F. Terminal year SSB and F were rho adjusted, as were starting numbers at age for short term projections. The rho adjustment amounted to scaling terminal year SSB and starting numbers at age by 0.53. There was no unusually strong year class in the starting numbers at age, so this rho adjustment was assumed to be adequate to apply to all ages in year *T+1*.

- 2019: This assessment used data through 2018. Stock status was not overfished and overfishing was not occurring (SSB>>SSB_MSY). The retrospective pattern decreased relative to the 2017 assessment update, with a rho of 0.7 for SSB and -0.44 on average F. Terminal year SSB and F, as well as initial numbers at age, were rho-adjusted. There were no issues with large estimated F on small year classes, as occurred in 2017, so the reported metric for fishing mortality was a simple average of ages 5-7 as was done in all assessments other than 2017. Terminal year SSB and F were rho adjusted, as were starting numbers at age for short term projections. The rho adjustment amounted to scaling terminal year SSB and starting numbers at age by 0.59. There was no unusually strong year class in the starting numbers at age, so this rho adjustment was assumed to be adequate to apply to all ages in year *T+1*.

Fisheries Management

Haddock have supported fisheries on Georges Bank for centuries, initially as bycatch, then as a primary target species. In the 1800s and early 1900s, most haddock were caught in the cod hand line and longline fisheries for fresh markets (Clark et al. 1982). With the development of frozen filet markets, ice machines and otter trawls, a directed haddock fishery developed in the 1920s (Jensen 1967). Distant-water fleets targeted haddock on Georges Bank in the 1960's and early 1970s but were excluded from U.S. waters by the 1976 U.S. Magnuson Act. An international boundary between U.S. and Canadian waters was established in 1984. In 2001, the eastern Georges Bank management unit was defined by the US-Canada transboundary management agreement (TMGC 2002).

In 1958, the U.S. Fish and Wildlife Service limited minimum trawl mesh to 4.5 in. (114 mm, Jensen 1967). The International Commission for the Northwest Atlantic Fisheries (ICNAF) regulated minimum mesh sizes as well as minimum fish sizes, spawning closures and annual quotas for haddock in the 1950s-1970s (Kulka 2012). A fishery management plan was developed for cod, haddock and yellowtail flounder by the New England Fishery Management Council in 1977 (Wang and Rosenberg 1997).

Since 1985, U.S. fisheries for New England groundfish have been directly managed by the Multispecies Fishery Management Plan of the New England Fishery Management Council (NEFMC 1985). The initial management strategy was based on input controls (days at sea, size limits, gear restrictions, time/area closures) with year-round are closures on Georges Bank since 1994 and substantial increases in minimum mesh sizes in 1982 (5 1/8 in., 130 mm), 1983 (5.5

in., 140 mm), 1994 (6 in., 152 mm), 1999 (6 in., 152 mm, diamond mesh or 6.5 in., 165 mm, square mesh) and 2000 (6.5 in., 165 mm, for all trawls). The minimum legal size was decreased in response to slower growth of recent dominant yearclasses. The New England groundfish management system transitioned to an output control system (annual catch limits and catch shares) in 2010 (NEFMC 2009).

TOR 1: Review existing research efforts, data, and habitat information in the Gulf of Maine and Georges Bank, identify any findings relevant to influences of ecosystem conditions on haddock, and consider those findings, as appropriate, in addressing other TORs. For processes that the working group deems important and promising that are not currently feasible to consider quantitatively, describe next steps for development, testing, and review of quantitative relationships and how they could best inform assessments.

Georges Bank haddock habitat models

Ecosystem variables are important drivers of the spatial distribution of fish, therefore the working group developed and reviewed habitat models which might explain changes in the spatial distribution of haddock over time. The working group considered species distribution models based on machine learning methodologies (Friedland et al. 2020). In this version of the models, salinity variables were considered as static fields as opposed to dynamic, and zooplankton data were smoothed over five year time steps to extend the model training period from 1976-2019 and the estimated habitat was made for the same suite of years.

The estimated size of haddock habitat has changed over time, declining from the late 1970s into the 1990s (Figure 5a). Habitat scores in both seasons and stocks areas have increased over the last two decades and most dramatically in the last six years or so (coinciding with the strong 2013 year class). Habitat gradients, measured as the difference in habitat scores between Gulf of Maine (GOM) and Georges Bank (GB), are generally negative in the spring suggesting a gradient of higher habitat scores on GB than in the GOM (Figure 5b). The fall gradients tend to be opposite, with the positive gradient suggesting a gradient favoring the GOM. The net or annual gradients were generally positive, but strongly positive in the last six years of the time series. The extent of the haddock habitat in spring was approximately 60,000 km$_2$ over much of the timespan, but has apparently doubled in recent years to over 120,000 km$_2$ (Figure 6). The fall habitat has historically been larger than the spring habitat at approximately 80,000 km$_2$, but has also increased, though not to the extent of the spring habitat, to approximately 100,000 km$_2$.

The occurrence distributions for haddock suggests habitat has increased recently with the increase in abundance, and there is likely greater overlap between spring and fall habitat distributions than previously observed. Though there is a habitat discontinuity between eastern and western GB, which may contribute to population separation for the respective areas, haddock habitat appears rather homogeneous over the entire range of the species.The distribution of habitat would be consistent with the hypothesis of a single stock in the GOM/GB region. The difference in seasonal habitat would support the hypothesis of seasonal movement of fish, although ideally this would be verified with direct measurements of movement (e.g., tagging data). The tendency for greater net gradients favoring potential movement from GB to GOM would be consistent with the hypothesis that recruitment on GB influences patterns of recruitment seen in the GOM.

There was concern among the WG that the spatial habitat scores were themselves driven by changes in spatial abundance of haddock, as opposed to habitat driving abundance. An investigation of the most important variables in the habitat model showed that the top three variables (March chlorophyll concentration, average spring distribution of Acartia spp., and December distribution of SST fronts) described a negative space for haddock (i.e., where haddock are not found), while the fourth most important variable (fall distribution of *Centropages typicus*) described the entire spatial field (Figure 7). Subtracting the first three variables from the fourth yielded the spatial distribution of haddock, although the working group could not derive a mechanistic explanation for how those variables directly affect haddock. Therefore, it was not clear whether a causal relationship between the habitat scores and the spatial distribution of haddock was established.

The working group also considered a minimum swept area abundance based on habitat-informed dynamic re-stratification for haddock, which may prove to be a useful alternative to current survey designs if survey coverage changes in the future. In the absence of a mechanistic understanding of the linkage between habitat scores and haddock distribution, the working suggested these research methods continue to be developed and explored in the future.

Georges Bank haddock spatiotemporal models

The working group reviewed another analysis which aimed to model spatial distribution of haddock on GB using Poisson-link delta model composed of encounter-probability and catch rates with a process structure of spatial, temporal and spatiotemporal components. Data collected in long time series of NMFS spring (1968-2019) and fall (1963-2019) bottom survey which encompassed contrast changes in haddock spatial distribution, biomass and bottom temperature were used in this analysis (Figures 8, 9, and 10). The role of climate change and fish density in affecting changes in haddock distribution was explored with Generalized Addictive Model (GAM).

Model predicted Area of Occupancy of haddock on GB varied substantially over time, with consistent trend between spring and fall (Figure 11). Easting and Northing shift were detected in both seasons in comparison to the 1960s, with a clear Northing shift in fall in the last 10 years (Figure 12). GAM model explained a high proportion of deviance in Area of occupy with 71.3% and 65.3% for spring and fall, respectively (Figure 13). Density-dependence made the greatest contribution to the variations in both seasons, which was consistent with the basin theory (MacCall 1990). Variability in east-west movement was related to fish density in both spring and fall (Figure 14), warming was most influential in shifting from south to north in fall (Figure 15).

Northing shift driven by recent warming in fall has important implication for haddock stock assessment, especially with continued warming from climate models in next 50 years in this region (Pershing et al 2021). For example in 2019 fall survey, more than 70% of biomass of haddock on GB was in stratum 29. The working group decided to include strata 29 and 30 in EGB assessment (ToR 3 of EGB haddock report). Density-dependent haddock expansion to the west supported the summary of haddock stock structure in ToR 12.

TOR 2: Estimate catch from all sources including landings and discards. Describe the spatial and temporal distribution of landings, discards, and fishing effort. Characterize the uncertainty in these sources of data.

Overview

Commercial targeting of haddock was minor until ice was taken on vessels to facilitate landing fresh haddock at the end of the 19th century (Jensen, 1964). Development of the otter trawl and steam trawlers at the beginning of the 20th century greatly increased fishing effort on haddock, and the introduction of filleting fish at the port in the 1920s ramped up market demand for haddock fillets (Jensen, 1964). Clark (1982) reconstructed haddock landings and catch at age going as far back as 1931. These data are included in the present assessment.

Digital databases for US landings, length samples, and age samples are available beginning in years 1963, 1969, and 1965, respectively. Total catch estimates for earlier years were from previous stock assessment reports (e.g., NEFSC 2019). US landings are reported by the statistical area where they were caught (Figure 16). Prior to 1994, port agents interviewed captains to assign landings to statistical area. Beginning in 1994, Vessel Trip Reports (VTRs) were required, and captains self-report the statistical area of fishing efforts where landings were caught. Statistical methodology was applied to allocate unknown landings to statistical area from 1994 to 2019 (Wigley et al. 2008a; Palmer 2008). There appear to be fairly clear delimitations in US landings between the statistical areas associated with the Georges Bank stock and the Gulf of Maine stock (Figure 17), aligning with the shelf of Georges Bank and the surrounding channels.

A program to put observers on US fishing vessels to record discard information was initiated in 1989. From 1989-present, total discards are estimated from the ratio of discarded haddock to kept of all species, where the ratio is calculated by year, quarter (or other suitable time step), gear and mesh type, and prorated to the total landings of all species in the same time-gear category to obtain total discards (mt) (Wigley et al. 2008b). Where samples for time steps within the year are sparse, imputation is carried out (i.e., a discard ratio from another sampling period is assumed). Prior to 1989, historic assessment reports describe incorporating discards that were expected for years with large cohorts; explicit details are lacking and those estimates are not reproducible.

Haddock has produced several extremely large year classes (1963, 2003, 2010, and 2013), and several other exceptional year classes (1975, 1978). These year classes tend to dominate landings, when of legal size, and dominate discards when undersized. Strong density-dependent growth is observed when these year classes are present in the population, and this manifests itself in the length frequencies, in the market categories, and at times in the management regulations to adjust minimum size limits in order to reduce discarding.

US Commercial landings

Gutted:whole weight conversion factor

Haddock are landed in gutted condition and a conversion factor is used to estimate whole weight. It is believed that the conversion factor for haddock originated in the 1930s. Brown (1965) mentions a more recent analysis of haddock that appeared to be consistent with the historic value of 1.14 gutted:round weight, and that potential seasonal differences due to gills being removed in certain months was expected to amount to a difference of 2% at most. Data from a pilot study to re-estimate a conversion factor was reviewed by the working group. Sampling was limited spatially and temporally, and the working group recommended pursuing this work with a balanced sampling strategy across gears, statistical area, and by quarter. It was also recommended to make note the spawning condition of the sampled fish to eliminate that as a source of variability. The previously assumed value of 1.14 was used for this assessment.

Length-weight relationship

Landed fish are sampled for length, and a subsample of those same fish have an otolith removed for ageing. No individual weights are taken, but port samplers record an estimate of the sample weights and corresponding length frequency of fish sampled. Expanding length frequencies of the sample to the total catch requires a length-weight equation. For a given length weight equation, one can compare predicted sample weight with observed sample weight as a check that the length weight equation is appropriate (see Legault and Blaylock, 2008). A seasonal length-weight equation for spring and for fall was derived from the NEFSC Bottom Trawl Survey, which began recording individual fish weights in 1992. Differences in gear and mesh between the survey and commercial vessels contributes to differences in observed length frequencies (Figure 18). Comparing annual length frequencies, the observed length range had the most overlap between survey and commercial landings in 2005 (Figure 19). An analysis comparing estimated port sample weights to predicted sample weights using seasonal length-weight equations from 2005 was made. In addition, a length-weight equation from Wigley et al. (2003), a re-estimation of Wigley et al. (2003), and a length-weight equation using all years, were used to predict sample weights (Figure 20. The current length weight equation, which was estimated for year 2005, performed the best (Figure 21), with differences < +/-10% in all years (Figure 22). The total number of samples weighed in a given year is variable (Figure 23), and given the preponderance of recorded port sample weights where the final digit, or final two digits, are zero (Figure 24), the precision of those weights is not expected to be high, and as such, small differences may result from rounding error (Figure 22). The current length-weight equation, from 2005, was used (spring a,b parameters: 0.0000060658, 3.10782; fall a,b parameters: 0.0000071186, 3.08054).

Landings trends

Landings data are summarized for the year range where digital records are available (1964-present). Annual landings peaked at over 50,000 mt in the 1960s, dropped precipitously in the 1970s, followed by a brief peak of 15,000-20,000 mt per year in the early 1980s (Table 1; Figure 25). Landings reached an all-time low in the 1990s, with the stock being declared collapsed in 1992. Rebuilding occurred in the late 1990s, and since then annual landings have been around 2,000-6,000 mt. The primary gear has been otter trawlers, with very minor contributions from gillnet and hook/line gears (Figure 26).

Haddock are landed in all months, but quarter 2 (April - June) has typically had the most landings (Figure 27). Market categories include Large, Scrod, and Snapper–this order reflects decreasing size of fish. The fishery used to land primarily Large and Scrod, but with the three strong year classes in the last two decades, Snapper haddock have come to reflect a substantial portion of landings (Figure 28). The pattern in landings by statistical area has changed through time (Figure 29). Before the mid-1990s, a large fraction of landings were reported from the statistical areas in eastern Georges Bank (561, 562), as well as western Georges Bank (521, 522, 525, 526). Year-round closed areas were introduced in 1994 to help rebuild the collapsed haddock stock, with a large portion of eastern Georges Bank falling in to Closed Area II. This halved the fraction of landings coming from eastern Georges Bank, with 521, 522, and 525 accounting for 75% or more of landings since then.

Mandatory reporting with VTRs began in 1994, and replaced port agent interviews as the means to collect area of landings and fishing effort. Mandatory dealer reports are the basis for total landings, and when combined with VTR information on area of landing allows for the apportionment of total landings to statistical area (and ultimately, stock area). A direct match between dealer landings and VTRs is ideal (level 'A'), but when a perfect match is not possible then an algorithm assigns a statistical area based on trip characteristics (Wigley et al., 2008). Levels 'A' and 'B' use vessel-related data (vessel permit, gear group, main species group, and month), while levels 'C' and 'D' reflect fleet-related data (ton class, port group, gear group, main species group, and calendar quarter) (Wigley et al., 2008). Approximately 75% of trips are matched at the 'A' level, and another ~15-20% at the B level (Figure 30). The fraction matched at 'C' level has increased slightly in recent years. This matching procedure between VTR and Dealer reports is necessary because there is not a unique identifier on the VTR that could link directly to the Dealer report.

Port of landings has also changed through time, initially reflecting a broader distribution across the states of Maine, Massachusetts, and Rhode Island (Figures 31 and 32). A regulation prohibiting the landing of lobsters in Portland, Maine by groundfish vessels greatly reduced the landings of groundfish at that port, and the resulting distribution is concentrated in 3 Massachussetts ports (Boston, Chatham, and New Bedford). Tonnage class of vessels landing haddock has generally been class 3 and 4 (51-150 and 151-500 tons, respectively); Figure 33).

US Commercial landings biosampling

Biosampling of commercial landings (i.e., sampling of length and otoliths for age determination) are made by geographic region, quarter, market category, and gear. Minimum sampling criteria were agreed to at NAFO in 1974, with a recommendation for sampling density of 100 lengths per 200 mt, and age samples of one fish per centimeter length group (NAFO 1980). The Georges Bank haddock stock catch at age is calculated by market category and half year. A comparison of half-year versus quarterly catch at age calculation was made in 2008 during the last Georges Bank haddock benchmark (NEFSC 2008), but precision was poorer at the finer temporal scale. Catch at age is not broken out by gear due to the vast majority of landings coming from otter trawl gear, and broad similarity in length frequencies.

Length and age sampling for the time series that these data are recorded are shown in Figures 34 and 35. Length sampling has generally met or exceeded the NAFO threshold, with the exception of recent years for the Large market category. Not many fish are landed in this category and it is difficult for port samplers to encounter sufficient trips from which to take Large samples. Age samples are summarized using the same metric as lengths for ease of comparison. The same trend of meeting/exceeding the NAFO threshold in most years is observed, as is the drop off in achieving sufficient samples of Large market category haddock in recent years.

Age was determined by examination of the thin-section otoliths. Accuracy was assessed by re-aging a subset of samples from species reference collections. Reference collections were aged by consensus between at least two NEFSC readers and an age reader from the Department of Fisheries and Oceans Canada, as part of regular ageing exchanges between the two laboratories. NEFSC completes regular testing to provide measures of consistency and accuracy. For 28 haddock samples, including 2,611 otoliths tested, agreement was relatively high (97.2% agreement, CV=0.42; https://www.fisheries.noaa.gov/resource/data/accuracy-and-precision-fish-ages-northeast). Results from re-ageing subsamples taken from previously aged fish indicate that age determinations at both labs are reliable (e.g., Sutherland et al. 2007): two accuracy tests had 77% agreement (6.2% CV) and 96% agreement (0.6% CV), precision tests had 95% agreement (0.5% CV) at NEFSC and 92% agreement (1.0% CV) at DFO, and the inter-laboratory results were 86% agreement and 1.8% CV.

US Commercial landings samples at length and age

Length frequencies show clear shifts in the mode when large year classes enter the fishery (Figure 36). Landings ranged from about 40 cm to 80 cm until the early 2000s, but in recent years, landings are packed in the range of 40-60 cm as a result of strong density-dependent reductions in growth. The length distribution across statistical areas shows strong correspondence in the mode, suggesting homogeneous mixing across the bank (Figure 37). Length frequency by gear generally shows good agreement, recognizing that very small amounts of landings are sampled for gears other than otter trawl (Figure 38).

Sampled age frequencies of landings track the large year classes remarkably well (Figure 39). Earlier in the time series, fish were landed as young as 2 years of age, but in the last 20 years, fish are generally not landed below age 3. Similar to the distribution of lengths across statistical area, the distribution of ages shows strong correspondence and reinforces the evidence for homogeneous mixing (Figure 40). Age frequency by gear generally shows good agreement, recognizing that very small amounts of landings are sampled for gears other than otter trawl (Figure 41).

US Landings at age

Landings at age were re-calculated for this benchmark for years 1987-2019 (Table 2), specifically to allow for a recalculation of weights at age given the updates to Canadian landings. Length at age and weight at age show declining trends since about 2000 (Figures 42 and 43). Consistent tracking of strong year classes is apparent from age 3 or 4 through about age 8 to 10, depending on the year class (Figure @ref(fig:laa_number_gbhaddockYrFig)). Correlations in the landings at age, as estimated by bootstrap resampling, shows variation in pattern and magnitude among years, and also shows both positive and negative correlations (Figures 44, 45, 46, and 47).

US Commercial discards

Discards were re-estimated for the years that observer data are available (1989-2019) (Figure 48, Table 3), specifically to allow for a recalculation of weights at age given the updates to Canadian landings. Discards are estimated separately for eastern and western Georges Bank by half year and then summed to get total discards. Discards peaked in 1994, a period of strict landings regulations aimed at rebuilding haddock; discards were estimated at over 2,000 mt (43% of total catch, from US and Canada, that year). Discards peaked again in 2006 and 2007, in large part due to slowed growth from the 2003 year class leading to many undersized fish. An emergency rule was enacted to reduce the minimum size. Discards peaked again in 2013-2016 as the 2010 and 2013 year classes reached the fishery. The precision of discards was low at the start of the observer program (CV ranged from 0.27-1.6 from 1989-1998), but has improved substantially, with a CV of 20% or less in most years since then.

Large and small mesh otter trawl account for the bulk of discarded haddock, with a small amount coming from the separator trawl in recent years (the separator trawl is designed to retain haddock while allowing for cod to escape) (Table 4, Figure 49). The number of observed trips is intended to achieve a target precision in estimated discards. Number of observed trips, and the fraction of total trips that were observed, has varied through time, generally increasing since 2000 (Figures 50, 51, 52, and 53).

US Commercial discards biosampling

Samples of the lengths of discarded fish are made by observers, and these are expanded to the total discard amount by calculating weights of the sampled length frequency. The length distribution of discarded fish by half year is similar to that seen in the spring and fall NEFSC Bottom Trawl Survey (Figure 18). Length-weight relationships by year-season were used to expanded length frequencies of discarded fish, and the age length key from the corresponding year-season survey was used to estimate discards at age (Figure 54). Discards track the strong year classes beginning at age 1 or 2, and persist through legal sized ages (Table 5). Reasons given to observers for discarding include: regulations, poor quality, not brought on board (sometimes a gear issue), no perceived market value, or unspecified reason.

US Recreational catch

Estimates of recreationally landed haddock on Georges Bank were near zero in the MRFSS and MRIP data bases. The re-estimation following correction of the sampling frame (to mail based survey rather than land-based telephone) was still near zero in many years. The few non-zero estimates were *extremely* imprecise and the maximum magnitude approached that of the removals from a NEFSC Bottom Trawl Survey. No further consideration or analyses was pursued for recreational catch.

Canadian Commercial landings

Canadian landings were re-estimated for years 1987-2019 due to concerns about the length-weight equation. As with US landings, Canadian landings on eastern Georges Bank track large year classes (Figure 55). The fishery lands fish at ages 1 and 2, whereas these ages are discarded in the US fleet (Table 6). Since the mid-1990s, there are proportionally fewer fish landed below age 3.

For full details on canadian landings and discards, please refer to the appropriate section of the EGB haddock report. The Canadian fleet only operates on eastern Georges Bank, and therefore the description and details are identical.

Canadian Commercial discards

Discard estimates from the Canadian scallop fleet are greatest at age 0, and persist though about age 4 or 5 in small amounts (Table 7). Discards are greatest on the strong year classes, which track well through the 2003 year class at age 1 (in 2004). The magnitude of discards is remarkably small in years since 2005, particularly given the exceptional size of the 2010 and 2013 year classes (Figure 56).

Total catch-at-age and mean weight-at-age

Total catch at age in numbers (ages 1+), including the estimates from Clark (1982) back to 1931, range from about 30-40 million fish until 1965-1966, when catches of 202 and 136 million were taken (the 1963 year class at ages 2 and 3; Tables 8 and 9). Catch dropped quickly from that point, ranging from a few million in the 1980s and 1990s, to 15-25 million in the years where the other extremely large year classes were present.

The catch at age show little variability up to the 1960s, when the appearance of the 1963 year class is seen to dominate the fishery in each year of its life (Figure 57). Age 2 fish accounted for a large proportion of the catch through the early 1990s, but their representation in the catch at age has diminished. Catch primarily occurs on ages 3 and older, notwithstanding the discards.

Weights at age are calculated from a a numbers-weighted average of the individual time series that total catch is composed of. The time series of weight at age in the catch shows an inverse trend to the population biomass trends (see TOR3), reflecting strong density-dependence in growth (Figure 58).

For the years where landings and discards were re-estimated for both US and Canada (1989-2019), Canadian catch was 2.5 times greater than US catch on average (Figure 59). This peaked in the 1990s, where that ratio averaged 3.5 and peaked at almost 9 in 1996.

Sources of uncertainty

The following list summarizes sources of uncertainties and the working groups best guess at the likely magnitude of uncertainty (qualitatively summarized as minor, moderate, or major).

- minor uncertainty in statistical area landed from the allocation procedure
- minor uncertainty of fish landed in statistical areas at border with Gulf of Maine stock (Figure 60)
- uncertainty of fish landed in statistical areas on the Scotian Shelf (for 2002-2019, this varies from 2-19% of total catch on Georges Bank; Table 10 and Figure 61)

- unknown uncertainty of over- or underreporting of haddock landings; a recent court case documented a large seafood dealer (“the codfather”) who mislabeled other groundfish as haddock due to having excessive haddock quotas compared to restrictive quotas on the mislabeled fish
- minor uncertainty of gutted:whole weight conversion
- unknown uncertainty of probabilistic assignment of age from age-length keys for year classes that are adjacent to the exceptional 2003, 2010, and 2013 year classes

- unknown uncertainty of estimating catch weights given no individual fish weights for commercial landings (US)
- unknown uncertainty with derivation of Canadian catch weights
- unknown uncertainty in DFO landings or discards

TOR 3: Present the survey data being used in the assessment (e.g., indices of relative or absolute abundance, recruitment, state surveys, age-length data, etc.). Characterize the uncertainty in these sources of data.

Biology

Growth

For the purposes of calculating population biomass and spawning stock biomass (SSB), weights at age are needed. Over the time series of surveys conducted on Georges Bank, there are years where no fish of a given age were observed, and therefore there is no weight to associate with numbers of fish as predicted by the stock assessment model (Tables 11 and 12), and the number of fish observed at a given age is sometimes quite low. If growth were deterministic, or had low variability, one might be able to predict size at age where direct observations are missing, and then calculate weight from a length-weight equation. However, density-dependent changes in length at age for haddock on Georges Bank have been observed since the exceptional 1963 year class, and a single time-invariant von Bertalanffy growth curve is inadequate. There are no gaps in observed ages in the time series of catch, and therefore catch weights are used to calculate January-1 weights at age for the population following the Rivard (1982) approach. It is assumed that little additional growth occurs between January-1 and spawning (assumed to occur in March), and therefore the January-1 weights are used to calculate spawning stock biomass.

Maturity and condition

The working group reviewed a working paper on spawning phenology of haddock on Georges Bank and Gulf of Maine (see Wuenschel_TOR3_Spawning phenology of haddock.pdf uploaded to the portal). This work identified that the NEFSC Bottom Trawl Survey in spring typically occurs during spawning season and captures fish in all spawning stages (developing, ripe, spent, and resting). The NEFSC fall survey occurs in the non-spawning season, capturing mostly resting fish that are transitioning to early developing. Spawning condition was found to be related to bottom temperature, with more post-spawning fish collected at higher temperatures recently (after accounting for shifts in survey timing). Overall, it appears that spawning may have occurred earlier, and ended earlier, in the period since 2010. A laboratory study by Trippel and Neil (2004) measured female haddock weight loss of 24.1% post-spawning. Consequently, comparing fish condition over the time series where the proportion of fish in each maturity stage differs is problematic. A further complication of condition measures such as Fulton's K (ref) and relative condition (Kn) (Le Cren 1951) was noted in Blackwell et al. (2000), who demonstrate that both of these measures are sensitive to the length classes analyzed, and that trends will be detected even when the underlying length-weight equation has not changed. Measures of fish condition for haddock are susceptible to all of these issues, and therefore observed trends likely reflect the combined effect of differences in spawning timing and annual changes in the length frequency of fish included in the calculation (Figure 62).

Maturity at age ogives were estimated during the GARM-III (NEFSC 2008), where annual, moving average, and a single ogive (over all years) were compared. It was determined then that a single ogive, using data from all years, was most appropriate for Georges Bank haddock. With each assessment update since then, the new maturity data was compared with the existing ogive to see if it differed significantly (Figure 63)). The time series ogive was kept through the 2012 update assessment, but new maturity ogives were used for 2011-2014 (added for the 2015 assessment update), 2015-2016 (added for the 2017 assessment update), and 2017-2018 (added for the 2019 assessment update). These new estimates were "event driven" in that only the new years of data were looked at and compared to the long time series of data already on the books. For this benchmark, the full time series was analyzed and annual estimates of a50 and slope were compared (Figure 64), as were the annual fitted ogives (Figure 65). For most of the time series, the annual a50 confidence intervals are wide (reflecting sample sizes) and generally overlap the time series mean and there were several years where annual estimates did not converge. At the end of the time series, the last 4 estimates of a50 are significantly larger than the time series mean. The parameter controlling the slope of the ogive was significantly below the time series mean in 3 out of the last 4 years (and at several other points in the years since 2000). Examining the annually fitted ogives, the last 3-4 years support a difference in maturity from earlier years. When combined with the trends in a50 and slope, it was decided to use two separate maturity ogives, one for years 1970-2015, and one through years 2016-2019 (Figure 66). Inclusion of 2016 was justified by the fact that the 2013 year class was 3 years old in 2016 and given it's magnitude and suppressed growth, it made more sense to include that year in the stanza that reflected the impacts of density-dependence. Maturity data prior to 1970 are not available, however a maturity ogive for 1931-1969 exists in the assessment files and that was retained.

Natural mortality

Natural mortality is assumed to be 0.2 at all ages and in all years. Data on maximum age observed in the available surveys and catch provide support for this assumption (Figure 67). The maximum age observed includes fish up to 19 years in the catch and in the surveys. In the US catch, the oldest ages have been observed in the last decade, while in the Canadian catch, the oldest ages were observed in the 1990s. The trends in maximum observed age over the time series matches the exploitation history of the fishery. Maximum age was high in the 1970s, dropped in the late 1980s and remained low for the US data sources (catch and survey) through the 1990s before steadily increasing to present. The Canadian catch and DFO survey still show maximum ages of 15-19 during the 1990s when the US declared the Georges Bank stock to be collapsed. This is likely due to the density-dependent distribution patterns showing a contraction in geographic range of haddock during periods of low abundance – fish density on eastern Georges Bank was maintained during this period and Canadian catches greatly exceeded those of the US fleet in these years (Figure 59). A simple calculation of exponential decline for a given instantaneous mortality rate yields a probability of observing a fish of a given age. The cumulative survival curves in (Figure 67) show that fish have a 5% chance of surviving to age 15 under a total mortality rate of 0.2. Recognizing that this stock has not been unexploited over the time series of observations, the fact that fish over age 15 have been observed provides strong support that M should not be expected to be higher than 0.2. The precision of age assignment has been high for the years where testing is recorded (https://apps-nefsc.fisheries.noaa.gov/fbp/QA-QC/hd-results.html). Rings on otoliths do not get appreciably closer together on older haddock, and the haddock age reader has confidence to at least age 20 (S. Sutherland, pers. comm.).

Distribution and stock structure

This topic is covered in TOR 12.

Survey Overview

NEFSC Survey Design

The Northeast Fisheries Science Center (NEFSC) operates two bottom trawl surveys (BTS), conducted in the spring ("S") and the autumn ("fall" or "F" in most figures), with coverage extending from the Bay of Fundy in the north to Cape Hatteras in the south. The spring BTS began in 1968, while the fall BTS began in 1963 (Grosslein 1969; Azarovitz 1981; NEFSC 1988). The survey follows a stratified random sampling design (SRS), using standardized sampling procedures and equipment and with between 350 to 400 stations sampled per seasonal survey in a given year (target of one station per 200 square nautical miles, NEFSC 1988). Survey strata are defined by depth zones, latitude, and historic fishing patterns (NEFSC 1988). Station allocation per stratum is based on the area of the stratum. In the case of small or narrow strata, typically those found on the edge of a shelf, a minimum of two tows are allocated so that a variance can be estimated–even if the area of that stratum would otherwise warrant only a single tow. The Georges Bank stock of haddock comprises 15 offshore strata (13-25, and 29-30; Figure 68), encompasing 19,164 km^2 (Table 13). The survey strata are defined by natural features rather than rectangles defined by latitude and longitude, as the commercial landings statistical areas, which creates some areas of non-overlap where the two types of polygons do not conform (Figure 69).

NEFSC Calibration for vessel and gear changes

The vessel and gear used to conduct the BTS has changed several times over the course of the time series, with efforts to standardize to the extent possible. The primary vessels for conducting the BTS were the R/V Albatross IV and the R/V Delaware II. A Yankee 36 trawl was used for the entire autumn survey (1963-2008) and for the spring survey (1968-1972, and 1982-2008). A Yankee 41 trawl with heavier BMV oval doors was used during spring 1973-1981 in an attempt to better sample pelagic fish such as herring (Azaravitz 1981). Portuguese-style polyvalent doors replaced the BMV doors in 1985 for both BTS surveys. Gear and vessel use, and calibration factors are given in Table 14. Calibration factors for doors (1.51 and 1.49 for calibrating weight and numbers, respectively) and vessels (0.79 and 0.82 for weight and numbers, respectively) have been estimated for haddock, but there is no conversion factor for the Yankee 41 trawl that was used in the spring during 1973-1981.

The NEFSC BTS transitioned to a new vessel, the R/V Henry B. Bigelow, in 2009. Paired tows with the Albatross IV took place in 2008 during the spring and fall BTS and also included site-specific tows in the summer (636 total paired-gear sets; NEFSC Vessel Calibration Working Group 2007; Miller et al. 2010). Initial calibration studies by Miller et al. (2010) looked at a variety of potential estimators of a constant calibration factor for numbers and weight that reflects the ratio of catchability of the Henry B. Bigelow to the Albatross IV. Prior to the Transboundary Resource Assessment Committee (TRAC) assessment in 2010, a subgroup of TRAC members met to review the application of the constant calibration factors. This working group focused on the three TRAC species: cod, haddock, and yellow-tail flounder. Length-based factors were considered, and the working group noted the paucity of data at the smallest length classes (Figure 70) and that the beta-binomial estimator and ratio estimator significantly diverged at these sizes. Advice from an external review panel on the Miller et al. (2010) calibration study recommended using the beta-binomial model instead of the ratio estimator of Henry B.

Bigelow:Albatross IV if the two methods were giving similar estimates, and also in cases where the ratio estimator obtained estimates that were greater than the beta-binomial model. For parsimony and to restrict estimates to a seemingly reasonable range, a segmented regression was fitted where the two estimators gave similar results, and the left endpoint was held constant for sizes below which the estimators diverged (18 cm for haddock). An upper end point was estimated as the point at which the regression would be held constant (50 cm for haddock). The lengh-based calibration was the same for both seasons, and esimates were made for numbers at length. No length-based calibration for weight was estimated. These calibration factors have been used since 2010 to scale the Henry B. Bigelow data to the Albatross IV time-series (i.e., given calibration factor ρ, Henry B. Bigelow data are divided by ρ to scale catches to the Albatross IV time series).

Miller (2013) examined 16 species in a hierarchical mixed effect framework that allowed for variation among and between paired tows. Previous analyses (Miller et al. 2010; Brooks et al. 2010) estimated calibration factors across all stations in the calibration study. Miller (2013) fit a range of conditional binomial and conditional beta-binomial models to paired tow data, and used AIC for model selection. The best performing model for haddock was a conditional binomial model with intercept and cubic spline smoother of size for both across-pairs and pair-specific random effects (Figure 71; see Table 2 in Miller (2013)). Two additional models had similar AIC scores that differed by 1.12 and 2.86 (Table 4 in Miller 2013), but did not give very different calibration estimates, and the model most similar to that considered in Brooks et al. (2010) had a substantially higher AIC (Figure 72). Model diagnostics for the best performing model showed good agreement with expected residual distribution (Figure 73). For this benchmark of haddock, the working group decided to proceed with the length-based calibration factors in Miller (2013). The new calibration spans all lengths, rather than holding the calibration factor constant for sizes <18 cm and >50 cm. Confidence intervals for the Miller (2013) calibration are wider than the current length-based calibration factors, and reflects the variability in pair-specific random effects in estimating the conditional relative catch efficiency at length (Figure 74). The uncertainty reflected in these confidence intervals is not currently incorporated into the estimated design-based variance for the spring and fall indices, and similarly, the uncertainty of the previous length-based calibrations was not incorporated into the variance estimate.

The impact of the Miller (2013) length-based calibration compared to the current Brooks et al. (2010) calibration is most discernible at the smallest sizes and in tracking the exceptional year classes (Figures 72 and 72). The total aggregate index in the fall has the largest difference, because it observes age 0 fish, and this age corresponds to lengths with the largest difference in length-based calibration (Figure 77).

Indices in number (and associated age composition in numbers) are used in the age-structured assessment models, but in the event that the assessment model is rejected, then the proposed “Plan B” approach would use biomass indices. Calibration of indices in biomass that account for the length based calibration at number had not been provided in previous estimation efforts. Biomass based calibration in Miller et al. (2010) scaled the constant calibration factors in number by the mean weight to obtain a constant calibration for biomass (0.878 in spring, 1.49 in fall). However, the mean weight during the calibration tows in 2008, relative to subsequent years, could be unrepresentative if the length composition were to differ. For haddock, where several extremely large year classes have moved through the population in the last 20 years, the length distribution and hence, the mean weight, have not been constant across years (Figure 78), and it is unlikely that the constant calibration for biomass would be appropriate in subsequent years (Brooks et al. 2010). A straight forward biomass calibration that accounts for this is to calculate the mean weight across lengths (*l*) per season (*s*) and per year (*y*) as

$$\bar{w}_{s,y} = \frac{\sum_{i=Lmin}^{Lmax} n_{i,s,y} a_{s,y} l_{i,s,y}^{b_{s,y}}}{\sum_{i=Lmin}^{Lmax} n_{i,s,y}}$$

where $a_{s,y}$ and $b_{s,y}$ are the season-year specific length-weight parameters, and $n_{i,s,y}$ is the number of fish at size *i*. From this, we can calculate the biomass calibration by season and year as

$$B_{s,y} = N_{s,y}\bar{w}_{s,y} = \sum_{i=Lmin}^{Lmax} n_{i,s,y} a_{s,y} l_{i,s,y}^{b_{s,y}}$$

Comparing these new season-by-year biomass calibration factors with the single seasonal biomass calibration from Miller et al. (2010) demonstrates the impact of the changes in mean weight (Figure 79). With the single seasonal calibration factor, the calibrated index is always greater than the uncalibrated index in spring because the calibration factor is (1/0.878), while the calibrated index is always less for the fall because the factor is (1/1.49). Using the new season-by-year biomass calibration factors results in a lower biomass index due to the combined effect of the new length-based calibration factor and the changes in mean weight due to annual length frequency (Figure 80). As a null check on the approach to calculate these new biomass calibration factors, the mean weight for uncalibrated Henry B. Bigelow data was multiplied by the uncalibrated N/tow index to obtain uncalibrated predicted Kg/tow. These predicted values were then compared to the observed Kg/tow (uncalibrated) with the expectation that the ratio should be close to 1. For this exercise, all but one value is between 0.96-1.03, and the lowest value was 0.92 (Figure 81).

Maintaining a single NEFSC time series versus breaking at the H.B. Bigelow years

A working group analysis explored the impact of breaking a contiguous index at the end of time series to understand the impact on model behavior (see Brooks_HaddockWG_Calibrate_Your_Expectations_March_29_2021.pdf uploaded to the portal). One concern was that splitting at the end of the time series would leave the model too uncertain about scale changes where the index is split, although others expressed confidence that the presence of the unsplit DFO survey would ease the model through this split. These explorations were performed by breaking the Albatross time series, which persists to 2008, at the year 1994, creating a "new" index for years 1995-2008. The current assessment model, a VPA, was fit to this new data, and a retrospective pattern emerged where none had been present before (rho on SSB went from 0.07 to 0.35). The VPA estimates of q at age for the new indices varied substantially through time for the "new" indices. The test was repeated by including 1 year of overlap between the original index and the "new" split index, and the retro improved slightly, but the q changes persisted. The analysis was repeated with ASAP to eliminate the chance that this result was a VPA indiosyncracy. The ASAP results had minor retrospective problems but q changed by +/- 20% and the estimated survey selectivities were different. A simulation in ASAP was also conducted so that the bias could be evaluated; scale differences were observed that related directly to the changes in q and selectivity. Based on these analyses, which showed poor performance when splitting the NEFSC indices (prior to the H.B. Bigelow vessel change) *and* with a contiguous index present (DFO), the working group agreed not to treat the NEFSC indices as a single time series for the purposes of assessment modeling, with the Miller (2013) calibration factors applied.

DFO Survey Design

The Canadian Department of Fisheries and Oceans (DFO) has conducted an annual bottom trawl survey on Georges Bank (statistical areas 5Zj, 5Zm, 5Zh, 5Zn, 5Zg, and 5Zo; Figure 82) between the 6th week of the year (Mid-February) until the 13th week (end of March), since 1987 (Figure 83). Due to vessel mechanical challenges the survey has been conducted later in a number of recent years (2015, 2107 and 2018). The survey follows a stratified random design for tow location. The stratum boundaries are generally based on depth contours as well as the location of the international boundary and geographic regions on the bank. Stratum depth ranges are < 93 m (50 fm), 0-183 m (0-100 fm), and 93-183 m (51-100 fm).

DFO Vessels and Gear

The CCGS Alfred Needler is the standard vessel used for the DFO Georges Bank survey, but when unavailable, the CCGS Wilfred Templeman, a sister ship to the Needler, was used in 1993, 2004, 2007 and 2008. In 2016 and 2017, the CCGS Teleost was used and in 2018 the Mersey Venture (a sister ship to the Teleost and an industry vessel) was used instead of the DFO survey vessel for the DFO Georges Bank survey. No conversion factors are available for the Templeman, Teleost or Venture, however, these vessels are considered to be similar in fishing strength to the Needler and was outfitted with the same gear. The standard DFO survey gear has been used for all years since 1987, relying on a Western IIA net with a 20 mm stretched mesh cod-end liner, using standard protocol.

Indices for the assessment model

Summaries of the data for indices proposed for inclusion in age-based assessment models for Georges Bank haddock are described below. This includes the two NEFSC BTS and the DFO BTS (in years where sampling occurred in all strata). The NEFSC Bottom Longline Survey (BLS) occurs primarily in the Gulf of Maine, and only part of one stratum (29) in the defined Georges Bank survey strata (Figure 84) is covered by the BLS. Therefore, the BLS is completely unrepresentative for the Georges Bank haddock stock. The Massachusetts Department of Marine and Fisheries (MADMF) BTS was also not appropriate for use on Georges Bank due to only covering inshore waters and the majority of haddock, in the limited instances when they are caught inshore, primarily occur in Cape Cod Bay and in the vicinity of Cape Ann (Massachusetts) (Figure 85). An index of landings per unit effort (LPUE) was briefly presented, but the trends bore absolutely no resemblance to other survey biomass trends; moreover, there were general concerns about how the change from effort management (trip limits, days at sea limits, etc.) for fishing years 1994-2009 to management by quota allocation for years 2010 to present was being handled in the index, as well as the lack of inclusion of discards. There was unanimous agreement to not consider the LPUE index for Georges Bank haddock.

NEFSC Survey Data

Haddock abundance and biomass were high at the beginning of the fall (1963) survey time series as a result of an exceptional 1963 year class (Figure 87). By the time the spring survey began (1968), most of that cohort had been fished out of the population. Over the next 30 years, population increases were driven by stronger than average year classes in 1975, 1978, and 1980, 1985, 1998, and 2000. In the last 20 years, biomass reached the highest levels ever observed, which was driven by 3 boomer year classes: 2003, 2010, and 2013.

There is a seasonal distribution of haddock where fish move to cooler, deeper waters in the fall and are more often found across the top of the bank in strata 21-25 and 29-30, and in stratum 16; in spring, survey timing often coincides with spawning and haddock are found on the Northeast peak of Georges Bank in stratum 16 and portions of 21-22 as well as spreading towards the southwest through stratum 13 (Figures 88 and 89,). In recent years, as abundance has increased dramatically, shallow strata in the center of the bank that are normally depauperate of haddock, accounted for 25% to 70% of the mean index in numbers in spring. Recent distribution changes were also observed in the fall, particularly in 2019 where stratum 29 accounted for 68% of the overall mean index that year (typical contribution is 10-20%). These distributional shifts can be more easily visualized from spatial maps of density, showing seasonal concentrations on different parts of the bank as well as expansions to more continuous distributions as population biomass increases (Figure 90 and 91 VAST fitted model, Brooks unpublished). This pattern appears to be a classic example of MacCall's Basin Model (MacCall 1990). A seasonal Gini index (Lorenz ref) similarly shows more concentration during very low abundance years and less concentration in high abundance years (Figure 92; after Wigley 1996).

In spite of the expansion and contraction associated with population density, the size and length distribution of sampled fish are similar among statistical areas (Figures 93, 94, 95, and 96). The fall survey encounters young of the year that were spawned the previous spring, but otherwise the surveys encounter the same range of sizes (Figure 97). In broad terms, smaller fish tend to be found at shallower depths, while larger fish (when present) tend to be in deeper waters in both seasons (Figures 98 and 99). There is a slight trend of larger fish being found at lower temperatures (Figures 100 and Figure 101).

Tow allocation to Georges Bank has ranged from 71-93 in the fall (with a few high sampling years in the 1970s), and 71-96 in the spring (Figure 102). The positive tow rate on these tows has reflected abundance trends of the population, ranging from 17-87% in the fall and 23-87% in the spring. Catch per tow histograms show occasional large tows, especially in recent years (Figures 103, 104, 105, and 106).

Sampling for the spring survey typically occurs from mid-March through April, and in recent years has extended into early May, while the fall survey typically samples Georges Bank from late September through the end of October (Figure 107). Sampled depths range from about 30 m to over 400 m, and average 120 m. Mean annual temperature in both surveys fluctuated over the time series, however the last ten years have been about 1 degree above the time series mean for both the spring and fall.

Haddock appear in the fall survey beginning at age 0 (i.e. fish that were spawned in the spring of the same year), and beginning at age 1 the spring survey. The strong year classes persist and are easy to track in both seasons (Figures 108, 109. Strong cohort tracking between ages is evident (Figures 110, and 111).

DFO Bottom Trawl Survey Data

The DFO survey time series was initiated in 1986, but that was considered a pilot year. In 1987, the survey sampled the whole of Georges Bank, with half of the tows in 5Z1-4 and half in 5Z5-8. Since then, tow allocation has been much higher in 5Z1-5Z4 (80-100%), the strata that define eastern Georges Bank (Tables 16 and 16). There are many years in the last two decades that the DFO survey did not sample the whole bank, and those years are not used in the assessment. The DFO survey averaged 83 tows on Georges Bank from 1987-2019, or 92 tows in years where the whole bank was sampled. Detailed annual Georges Bank Ecosystem Research Vessel (RV) Survey reports are available on the Canadian Science Advisory Secretariat website (Publications (isdm-gdsi.gc.ca); DFO 2019) and Stone and Gross (2012) provides a review of the Georges Bank Research Vessel survey program, 1987-2011.

Haddock appear in the DFO survey beginning at age 1. The strong year classes persist and are easy to track across years (Figure 112).

NEFSC and DFO Survey Trends

For the years that the 3 surveys overlap (1987-present), there is strong agreement in trend, showing a steady increase from the 1990s in both biomass and abundance, peaking in the mid-2010s, and some decline to 2019 (Figures 113 and 114). It is difficult to compare the indices for the many years in the DFO series that did not survey the entire bank, because the index in those years only sampled in the highest density strata on eastern Georges Bank. In mean numbers per tow, the time series maximum for both the NEFSC spring and fall occurs in 2014, but the maximum for DFO occurs in 2015 (only 47 tows were made, all in 5Z1-5Z4). Coefficients of variation (CV) for the indices tended to range between 0.2-0.4, with a few years earlier in the time series with CV of 0.6 (Figure ??).

Sources of uncertainty

The following list summarizes sources of uncertainties and the working groups best guess at the likely magnitude of uncertainty (qualitatively summarized as minor, moderate, major, or unknown).

- all calibration factors (gear, vessel) are applied without accounting for uncertainty of these estimates when estimating the variance of relative indices; no calibration for Yankee 41 trawl
- the influence of a single very large tow in two years of the Autumn and Spring NEFSC surveys is large
- not sure if DFO tows show similar issue of influential tows
- ‘minor?’ uncertainty among AL-HB calibration methods, particularly for small sizes
- no surveys in 2020
- moving forward DFO will no longer sample WGB

TOR 4: Estimate annual fishing mortality, recruitment and stock biomass (both total and spawning stock) for the time series, and estimate their uncertainty. Compare the time series of these estimates with those from the previously accepted assessment model, and evaluate the strength and direction of any retrospective pattern(s) in both the current and the previously accepted model. Enumerate possible sources of the retrospective patterns and characterize plausibility, if possible.

This TOR was addressed by updating the current model (Virtual Population Assessment, VPA) with new data to the extent possible, examining the impact of new data treatments on the scale of abundance and the measure of retrospective pattern (Mohn's rho). Next, a bridge was built between the VPA and a statistical catch at age model (Age Structured Assessment Program, ASAP; Legault and Restrepo 1998). This bridge was primarily a stepping stone en route to modeling the proposed base model in a state space framework (Woods Hole Assessment Model, WHAM; Stock and Miller 2020), although model structure, diagnostics, fleet selectivity blocking, and post-hoc weighting adjustments were examined for ASAP runs. A full exploration of model configuration and model building was pursued in arriving at the proposed base model in WHAM, as detailed below. All assessment models were fit to data from ages 1 to 9, where age 9 is a plus group. The decision about plus group age was justified based on identifying the oldest age that was seen with sufficient frequency in the surveys and the catch; beyond about age 9, there is are more zero observations than non-zero observations (Tables B11 and B12). Furthermore, growth tends to slow around age 7, asymptoting around age 9.

Data that are fitted in the models included catch at age and total catch in biomass from 1931, and three fishery independent indices of abundance (NEFSC Spring, NEFSC Fall, and DFO) that were fitted as age-independent indices in the VPA or as aggregate indices in number with associated age composition (ASAP and WHAM).

VPA

The VPA software (ADAPT, Gavaris 1998), can be obtained from the NOAA Fisheries Integrated Toolbox. The currently accepted VPA model for Georges Bank haddock fits to total catch at age (assumed to be known without error) and is tuned to 30 age-specific indices of abundance: NEFSC-Spring (Yankee 36 net) for ages 1-8; NEFSC-Spring (Yankee 41 net) for ages 1-8; NEFSC-Fall for "ages" 1-6, and DFO ages 1-8. The VPA assumes that indices occur at the beginning of the year, and therefore "ages" in the NEFSC-Fall survey are relabeled such that age *a* in year *y* is input as age *a+1* in year *y+1*. The model is tuned by including the observed index in year *T+1*, where year *T* is the terminal year of catch data. The objective function is the sum of squared log scale residuals for all 30 indices, with equal weight given to each index.

The last management track assessment was conducted in 2019 (NEFSC *in prep.*), with catch data through 2018 and indices through 2019. For this research track assessment, an additional year of catch data is available (2019) but no indices are available in 2020 (due to the Covid-19 pandemic, NEFSC surveys were not conducted on Georges Bank and the DFO survey did not cover all strata on Georges Bank); therefore, the VPA was not updated with the additional year of data, and all comparisons with working group data decisions were made with the VPA from the management track assessment.

There were two major data decisions made by the working group that altered the input to the VPA: using the Miller (2013) length-based calibration factor instead of the Brooks et al. (2010) factors; and the re-estimation of catch at age and weight at age for the Canadian fleet from 1987-present. Additionally, during the review of TOR 3 (Survey information), three years of the DFO survey with incomplete coverage (1993, 1994, 2010) had been included in the VPA as well as 1986 (considered a pilot year for the survey); those four years were dropped and the impact was evaluated. Lastly, a decision was made during exploration of model structure in ASAP and WHAM to drop the NEFSC-Spring survey years 1973-1981, corresponding to the years that a different net was used and no calibration was available. A sensitivity run with those years dropped from the VPA was also explored. The impact of the two major data decisions (new calibration factor, and new catch and weight at age) were explored separately and in combination, while the two minor adjustments (drop four years from DFO survey, and drop Yankee 41 years from Spring Survey) were explored as additional changes to the model with both major changes.

Changing the calibration factor reduced the value of Mohn's rho on spawning stock biomass (SSB) from 0.699 to 0.617, while replacing the catch at age for the Canadian fleet produced a larger reduction in Mohn's rho on SSB to 0.576; the combined effect of these two changes reduced the retrospective pattern to 0.492. Dropping four years from the DFO survey, given the new calibration and catch at age, had a minor impact and rho on SSB was reduced from 0.492 to 0.468 (Table B17). With those 3 changes in the VPA, dropping the Yankee 41 years from the NEFSC Spring survey had no further impact on the rho for SSB. The impact on estimated SSB in 2018, after applying a rho-adjustment, was a reduction from 505,937 mt to 454,965 mt (95% bootstrap interval of 309,443-664,725 mt).

The value of Mohn's rho for average F (over ages 5 to 7) was -0.441 in the 2019 Management Track Assessment. There was almost no change with the new calibration factors (-0.447), but rho was slightly worse with the new catch at age for the Canadian fleet (-0.451). When both of the data changes were made to the VPA, the rho on average F was -0.452. Dropping the four years of the DFO survey and then dropping the Yankee 41 years from the NEFSC Spring survey both resulted in a rho of -0.431 (i.e., no change when the Yankee 41 years were dropped, given all other changes in the VPA).

The impact of the data changes on the full time series of VPA results for the average F (reported as average over ages 5-7), the fully selected (or "apical") F, recruitment, and SSB are shown in (Figure 116). Differences are first observed in the time series where the new catch and weights at age occur (1987), and the next point at which differences are observable occurs for the years where the new calibration factors were applied (2009-2018). The new catch at age resulted in slightly larger estimates of recruitment, minor differences (both positive and negative) in fishing mortality estimates, and a reduction in SSB. The new calibration resulted in smaller estimates of recruitment, reduced SSB, and very minor differences in fishing mortality. The combined effect of both of these changes (and the minor changes of dropping DFO years and Y41 years of data) was reduced SSB estimates, and fluctuations in the recruitment and fishing mortality estimates.

Estimates of catchability at age for the indices were stable, except for models that dropped three years of DFO data — in those models, catchability for ages 5-8 increased (Figure 117). Implicit fishery selectivity at age (calculated as the fishing mortality at age in a given year divided by the maximum fishing mortality for a given year) shows generally increasing selectivity to age 7 (Figure 118); selectivity at age 8 is estimated in the model as the average fishing mortality at ages 5-7, and the F on the plus group is set equal to the F on age 8 (the oldest "true" age). In many years, age 5 is not fully selected, so including it in the average has the effect of introducing doming on ages 8-9. A sensitivity to the base model where the average F for ages 6-7 was used to specify F for ages 8-9 showed slightly higher selectivity estimates for ages 8-9 (Figure 119), but it was not possible to explore a completely asymptotic selectivity configuration due to the lack of separability between F and selectivity in the VPA.

Building the bridge to ASAP

The Age Structured Assessment Program, ASAP (Legault and Restrepo 1998) software, can be obtained from the NOAA Fisheries Integrated Toolbox. As described at the NFT software website, ASAP is an age-structured model that uses forward computations assuming separability of fishing mortality into year and age components to estimate population sizes given observed catches, catch-at-age, and indices of abundance. The separability assumption is partially relaxed by allowing for fleet-specific computations and by allowing the selectivity at age to change in blocks of years. Weights are input for different components of the objective function which allows for configurations ranging from relatively simple age-structured production models to fully parameterized statistical catch at age models. The objective function is the sum of the negative log-likelihood of the fit to various model components. Catch and survey age composition are modeled assuming a multinomial distribution (given fixed input weights for effective sample size), while most other model components are assumed to have lognormal error. Specifically, lognormal error is assumed for: total catch in weight by fleet, aggregate survey indices, stock recruit relationship, and annual deviations in fishing mortality. Recruitment deviations are also assumed to follow a lognormal distribution, with annual deviations estimated as a bounded vector to force them to sum to zero (this centers the predictions on the expected stock recruit relationship). The variance for each of these lognormal components is specified by the user (input as a year-specific or overall CV). For more technical details, the reader is referred to the technical manual (Legault 2008).

The first step in building a bridge to the ASAP model was to configure ASAP as closely as possible to the VPA model. Data from the VPA were directly imported, and the default ASAP settings were adjusted to more closely match the way that the VPA estimation is done. Specifically, a CV on aggregate catch was fixed at 0.03 and the effective sample size on catch age composition was set at 150 – this was an attempt to force ASAP to match the catch nearly exactly (VPA assumes catch is known without error). Indices were treated exactly as they are in the VPA, and were input as individual indices-at-age with a catchability at age estimated for each index; the CV associated with each index was 0.4, consistent with the VPA treatment of all indices being equally weighted in the objective function. The Yankee41 spring series was not included in ASAP, consistent with the final update to the VPA. No stock recruitment function was estimated; instead, steepness was fixed at 1 and all recruitment was freely estimated. While the VPA does not have a separability assumption between fishing mortality and selectivity, ASAP does. A single block of constant selectivity for all model years was fit at age, with age 7 being assumed to be fully selected (based on VPA results suggesting this age was most frequently fully selected). This ASAP model (ASAP_Bridge1) had no retrospective pattern, and the 95% confidence interval around the terminal year SSB estimate overlapped with the VPA 95% bootstrapped rho-adjusted SSB in the terminal year (Figure 120; only the VPA SSB in 2018, the terminal year, is rho-adjusted). Recruitment estimates were noticeably smaller in the ASAP run, and the single fleet selectivity block in ASAP is likely the cause for the difference between average F on ages 5-7 in ASAP versus VPA estimates. The average VPA selectivity at age (averaged over 88 years) is quite similar to the single selectivity at age estimated in ASAP_Bridge1.

Two additional ASAP bridge runs were conducted. ASAP_Bridge2 simply updated the maturity ogive, and this had no noticeable impact on any model estimates. ASAP_Bridge3 estimated logistic fleet selectivity instead of selectivity at age, and it had similar SSB and recruitment estimates as the other ASAP bridge runs, while selectivity smoothed through the earlier "at-age" selectivity pattern (Figure 120). No further bridge building was pursued after ASAP_Bridge3. Full diagnostics for ASAP_Bridge2 and ASAP_Bridge3 are in the zipped ASAP folder (**ASAP.zip**). The maximum gradient for both run exceeded 1E-3, and the age composition residuals for the catch showed non-random patterns and was considerably worse for ASAP_Bridge3.

ASAP further model exploration

The next ASAP modeling explorations followed a more standard structure for statistical catch at age models, with indices input as an aggregate index (in numbers) with age composition data, and user-specified fixed CV on the aggregate indices and user-specified fixed effective sample size. Common to all of the ASAP models described below, are the following: no stock recruit function estimated (steepness fixed at 1); recruitment was estimated as a mean with annual deviations governed by a fixed input CV=1.0. Working group data decisions included were: the new maturity ogive was used; the new length-based calibration factors were used; the DFO years 1987 and 1993-1994, and 2010 were dropped. Additional differences, and key results are highlighted in the paragraph describing each of these ASAP exploratory runs. The CV on aggregate catch was set at 0.2 for years 1931-1962 and 0.1 for years 1963 to the final year estimated. The motivation for this is that there are no electronic records to be able to recreate the total catch or the catch at age back in time; this setting gives the model more flexibility to fit those historic catch data. Similarly, an initial guess at the fixed effective sample size for catch at age was 50 for years 1931-1962 and 70 for years 1963 onwards, again expressing slightly less certainty in the data that is unavailable in electronic databases. The CV for aggregate indices was set to annually estimated CV from the simple random sampling calculation. Effective sample size for all 3 indices was initially set at 40 for all years that age composition data exist (40 is roughly half of the total number of tows in a given year). There was a desire to use indices on a relative scale (mean N/tow from the stratified random sampling calculation) because of the various spatial models that were being explored in the working group, with the understanding that q estimates could be scaled up to the area of whatever spatial unit was being explored *post hoc*. A single run in the first exploratory model (ASAP_1Block) made a comparison of results when indices were input as swept area or on a relative scale, and differences were negligible; thus, all models described below used aggregate indices in units of mean N/tow. A key difference with these ASAP models and the previous bridge runs is that the indices are now input for the year and age that they truly correspond to (and all ages are represented, including the plus group), and are matched to the month when the surveys take place, as opposed to the shift of 1 age and 1 year to accommodate the assumed January-1 start in the VPA (which uses only ages 1-8 or "1"-"6"). Age-specific selectivity was estimated for the indices and the fleet, requiring that at least 1 age be fixed at full selectivity. Those initial fixed ages were: ages 4-7 for the fleet; age 2 for the NEFSC spring index; age 1 for the NEFSC fall index; age 3 for the DFO index. This age range was determined initially from catch curves to find the "peak age" of full selectivity, and then by inspection from the ASAP output to make sure that no ages hit the boundary – in the case of the fleet selectivity, age 4 was initially fixed and subsequently ages 5-7 were also fixed at 1 due to boundary solutions.

A final configuration common to all ASAP models below is that the NEFSC Spring index using the Yankee 41 net was dropped. The index consisted of only 8 years, and estimating selectivity at age led to nearly all ages being estimated at the upper bound of 1.0. Attempting to estimate logistic selectivity, which required estimation of only 2 instead of 8 selectivity parameters, resulted in convergence issues as the model attempted to estimate essentially a flat line at 1. For the ASAP models, it was decided to drop this index due to the limited number of observations and inability of the model to estimate selectivity.

Model building in ASAP requires multiple iterations of looking at diagnostics, fixing parameters (if any) that are estimated at a boundary, testing selectivity shapes (selectivity at age, which can accommodate non-smooth shapes, or logistic selectivity) for fleets and indices of abundance, and examining age composition residuals for runs of all positive or all negative residuals for a given age(s) across a number of consecutive years – such a pattern suggests exploring whether introduction of a new selectivity block (in the case of fitting to a fleet) or a different selectivity function is supported. When a final model structure has been decided, a final model step is a *post hoc* adjustment to the input CVs on indices (to bring the RMSE to ~1 for each index, depending on the number of observations for the series), and effective sample size is multiplied by a scalar as suggested in Francis (2011).

ASAP_1Block_2018: This run used data through 2018 for direct comparison with the bridge model ASAP_Bridge2. Results were very similar for SSB, average F, recruitment, and fleet selectivity (Figure 121). Index selectivities were somewhat different, generally reflecting the shifting that was done to age/year in the VPA, and also the fact that 9 ages were used for all indices. Diagnostics indicated a number of boundary solutions for selectivity parameters, trends in residuals for the age composition, and RMSE greater than the expected value of 1 (for indices), but no further tinkering was done on this model as it was solely for comparison of the impact of changing from individual indices at age to aggregate indices with age composition.

ASAP_1Block: An additional year of data was added to ASAP_1Block_2018. This model and all subsequent models used data through 2019. Diagnostics showed strong blocking of residuals for aggregate catch, catch age composition The retro for this model was small (0.06 for SSB, -0.1 for average F).

ASAP_2Block : Given the residual pattern in the 1Block model, a new selectivity block was introduced at the earliest run of residuals in 1950. With 2 selectivity blocks for the fleet in the model, diagnostics still showed strong blocking of residuals in the total catch and catch age composition. The retro for this model was small (0.04 for SSB, -0.06 for average F).

ASAP_3Block : A third selectivity block was introduced for the fleet in 1974. Residual patterns persisted, but were somewhat improved in the earlier part of the time series. The retro for this model was 0.18 for SSB and -0.21 for average F.

ASAP_4Block : A fourth selectivity block was introduced for the fleet in 1990. residual patterns were improved over the models with fewer selectivity blocks, but some blocking remained in the most recent years. The retro for this model was 0.24 for SSB and -0.24 for average F.

ASAP_4Block_ADJCV_NEFF : This model configuration was identical to the ASAP_4Block model, but post-hoc adjustment was made to the CV on aggregate indices and effective sample size for the age composition data (catch and indices). Effective sample sizes were roughly halved from their initial guesses. The SRS variance associated with each index was scaled by a constant for all years until the resulting RMSE minimally achieved the expected value given the number of index observations; these scalars increased the standard deviation by 14% for spring, 52% for fall, and 16% for DFO. Residual patterns looked better than all other models examined, and the retro was smaller than the 4 block model without post-hoc adjustment. The retrospective rho was 0.17 for SSB and -0.18 for average F.

ASAP_4Block_Flat_ADJCV_NEFF : This model has the same post-hoc adjustments as ASAP_4Block_ADJCV_NEFF, but fleet selectivity in all blocks is logistic. Due to post-hoc weighting, the AIC of these two ASAP models can't be compared to the models without post-hoc weighting, but can be compared to each other. AIC favored this logistic configuration, but the retrospective rho was 0.438 for SSB and -0.323 for average F.

ASAP_5Block : A fifth selectivity block in 2002 was introduced for the fleet. A year on either side of this was tested for improvement, but the objective function was lowest at 1990. Residual patterns were slightly improved, but the retro was worse at 0.34 for SSB and -0.248 for average F.

General conclusions from these ASAP models are that the addition of selectivity blocks for the fleet provided a better fit to the aggregate catch and catch at age (as justified by AIC and diagnostic plots), but the additional parameter estimates translated into less precise estimates and increased retrospective patterns. The post-hoc adjustments made to the 4 block model produced the most acceptable diagnostics. Full diagnostics of each of these runs is in the asap zipped file on the share drive (**ASAP.zip**). Selectivity functional form for the fleet (in all blocks) and the indices was age-specific, and resulted in unusually strong doming. Exploration of fits using logistic instead of age-specific forms was not pursued in ASAP but was in WHAM.

WHAM overview

The Woods Hole Assessment Model, WHAM (Stock and Miller 2021), is a state-space stock assessment model with similar structure to ASAP (fits are to aggregate indices and aggregate catch, each with associated age composition), but with many options for incorporating process error: in the transition of abundance from age *a* in year *y* to age *a+1* in year *y+1* ("survival"), directly in natural mortality (*M*), in the stock recruit relationship, in selectivity, and in catchability (*q*). *M*, recruitment, and *q* can be linked with an evironmental correlate. Process error can be modeled as independent and identically distributed deviates (iid) or as an auto-regressive process (AR1 with correlation across ages/parameters or across years, or 2DAR1 across both ages/parameters and years). As discussed in Stock and Miller (2021), random effects in the abundance at age transitions can be due to variability in survival (natural mortality, e.g.) but can also reflect immigration/emigration, misreported catch (landings and/or discards), or misspecification of selectivity of the fishery or indices.

Aggregate indices and catch are assumed to have lognormal error distribution, with fixed input standard deviation, similar to ASAP, although it is possible to attempt estimation of a scalar for those observation errors. Whereas ASAP models use the multinomial distribution for fitting age composition data, WHAM currently includes 5 different options for modeling age composition data (multinomial, Dirichlet-multinomial, Dirichlet, logistic normal that treats zero values as missing, or logistic normal that pools zero values). The last 3 likelihoods are considered 'self-weighting' because they estimate a dispersion parameter that scales the input effective sample size. In addition, the logistic normal and Dirichlet distributions allow for both positive and negative correlations, which has been argued to be more reflective of the correlations observed in the data (Francis 2014). WHAM can be downloaded from the NOAA Fisheries Integrated Toolbox or from the WHAM Github repository, where a full suite of vignettes illustrates many of the capabilities of WHAM.

WHAM exploration part I – "WHAM as ASAP"

The first transition step from ASAP to WHAM involved importing the ASAP_1Block model into WHAM, and then fitting a non-state-space ASAP-like model to demonstrate the capability of WHAM to reproduce ASAP results when configured with similar assumptions (and specifying multinomial for the age composition likelihood). The only structural difference with this exploration is the way that recruitment is estimated. In ASAP, a mean recruitment was estimated with annual deviations from that mean that are controlled by user-specified fixed CV on those deviations. In WHAM configured as ASAP, recruitment in each year is estimated as a fixed effect parameter. In all cases examined here and in other applications, this has produced negligible differences in results between ASAP and WHAM model fits.

All ASAP models with 1-5 blocks for fleet selectivity were fit in WHAM with both age-specific selectivity and logistic selectivity (Table B18). Similar to ASAP results, adding more blocks to the fleet allowed the model to fit the catch data better, resulting in lower AIC, but also creating larger retrospective patterns. The other important pattern was that age-specific selectivity produced doming, and consequently, biomass estimates were lower with logistic selectivity while average F was slightly higher. Also, biomass was lower when additional blocks were added to the model. Full diagnostics comparing age-specific and logistic models for 1-5 blocks are in the zip file **compare_ASAP_fleetblocks_wFlat**, and model-specific diagnostics, including plots showing how parameters change over the retro peels, are included as individual zip files with names "WHAM_ASAP_1Block", "WHAM_ASAP_1Block_flat", …, "WHAM_ASAP_5Block_flat".

WHAM exploration part II – kicking the tires, one at a time

Exploration of WHAM model structures then proceeded by fitting indices and fleet selectivity at age, and exploring the impacts of just the different age composition likelihoods on results, with no random effects in the model. Next, random effects were incrementally added to the model. During this exploratory phase, the NEFSC Spring Yankee 41 time series was included in the model. Whereas it was dropped in ASAP due to insufficient observations to estimate separate selectivity and catchability parameters, in WHAM I explored the possibility of linking the selectivity to the full time series of NEFSC spring and estimating a separate catchability parameter. This was motivated by the description in Azaravitz (1981) that the Yankee 41 Trawl opened 2 meters higher than the Yankee 36, but much of the remaining gear configuration was similar (Table B19).

Results from exploring different age composition likelihoods with no random effects in the model were that the Dirichlet-multinomial model did not converge, the Dirichlet model had more boundary solutions (selectivity estimated at 1) than the other models, and the fleet selectivity was approximately asymptotic for the logistic-normal but domed at the oldest age for multinomial and Dirichlet. The differences observed in selectivity for the age composition models was also reflected in the estimates of catchability, with slightly larger q estimates for the logistic-normal likelihoods.

Next, random effects were introduced in the numbers at age (NAA), with a separate σ estimated for recruitment, and for ages 2-9+. All possible random effect options were tested: iid, AR1 correlation by age, AR1 correlation by year, and 2DAR1 correlation by age and year. These configurations were explored for all 5 age composition models. Again, many Dirichlet-multinomial models did not converge. Across the remaining age composition models, both 2DAR1 and AR1 correlation structure consistently performed best by AIC score. As in the prior step (with no random effects anywhere in the model), fleet selectivity continued to look asymptotic.

Building on models with random effects in NAA, random effects were introduced into selectivity, which was still modeled as age-specific for both fleets and indices. Similar convergence and model selection patterns persisted: dirichlet-multinomial had many convergence problems, 2DAR1 random effects were strongly favored over AR1 for both logistic-normal age composition likelihoods, while AR1 structure (across years) tended to have more support for the Dirichlet and multinomial likelihoods. There was no support for iid or AR1 by age random effects in selectivity. Fleet selectivity for the best performing models in each age composition likelihood tended to look asymptotic, particularly for the last several decades.

Several decisions were made based on this modeling exercise. The working group decided to drop the Yankee 41 survey. There were no convergence issues for this model, but the working group recommended dropping the Yankee 41 index for two reasons: i) it was short (only 8 years) and a ‘nuisance’ for estimation; ii) calibration information is not available to link the Yankee 41 to the Yankee 36 as a single time series. Therefore the Yankee 41 survey data were dropped from the model and not considered for further exploration. Also, given that fleet selectivity was tending towards asymptotic form, the working group recommended that the next modeling steps should test logistic selectivity for the fleet and potentially for the NEFSC spring and the DFO survey. No decision was made at this point with respect to the most appropriate age composition likelihood to use, and it was acknowledged that AIC could not be used for this decision.

WHAM exploration part III - reducing the factors

This exploratory phase proceeded in three steps with factorial model fitting. For all steps, model structures were explored across four age composition likelihoods (Dirichlet-multinomial had many convergence issues up to this point and was no longer explored). There were 10 model structures for incorporating random effects: 2 models for recruitment only (iid, AR1 across years), 4 models for random effects in all NAA, and 4 models for random effects in both NAA and fleet selectivity (iid, AR1 across years or ages/parameters, and 2DAR1). Thus a total of 40 models for the combinations of random effects and age composition likelihood. From this common set of model structures, specifics for the three steps were:

- Step 1: The ASAP 4Block model structure was explored, to allow the possibility of non-asymptotic fleet selectivity earlier in the time series. The time period for blocks were (1931-1949), (1950-1973), (1974-1993), and (1994-2019), with random effects just estimated for the last time block. Both fleet and index selectivity were age-specific.
- Step 2: A single selectivity block for the fleet was specified and random effects were estimated for all years.
- Step 3: A single selectivity block for the fleet, as in Step 2, but logistic selectivity was estimated for the fleet and for the NEFSC spring and DFO surveys, while age-specific selectivity was estimated for the NEFSC fall index.

Observations from this exploration were that in non-converged models, often a selectivity parameter hitting the upper bound of 1 was the culprit. Within a given age composition likelihood, models with random effects in both NAA and selectivity were preferred, typically 2DAR1, and fleet logistic selectivity also was preferred. Models with age-specific selectivity for the indices were preferred, but it required a lot of manual interaction to fix boundary solutions. There were no visible differences in model diagnostics looking across the different age composition likelihoods, but it was noted that the logistic-normal versions (either pooling zeros or treating them as missing) had slightly more converged models than the multinomial or the Dirichlet.

Literature on modeling age composition was reviewed to make progress on selecting an age composition likelihood in order to move forward. With respect to the multinomial, it requires the user specifying a fixed effective sample size for each year. Pennington and Volstad (1994) examined survey data of mean length for Georges Bank haddock. They calculated effective sample size as the number of fish that would need to be randomly sampled from a population in order to obtain the same precision as that obtained from the fish sampled on a survey. By their calculations, effective sample size for mean length was substantially less than the number of tows due to intra-haul correlation, and similar calculations could be carried out for age frequency distributions. Francis (2014) proposed that the Dirichlet and logistic normal would be improvements over the multinomial because they are self-weighting, that is, weights are estimated in the assessment model rather than the user fixing them *a priori*. For this reason alone, those likelihoods should be preferred over the multinomial. Beyond that, Francis (2014) noted that the logistic-normal allows for both positive and negative correlation among ages (which he demonstrated was observed, and even expected, in real data), while the Dirichlet (and multinomial) only allow for negative correlation. With respect to pooling zeros, Francis (2014) advised to pool only if the zeros were randomly distributed. A recent simulation study by Fisch et al. (2021) suggested that when sample size of age composition is moderate to large and process error is at least moderate, then the logistic-normal was a reasonable choice; however, if composition or sample size was small, process error was negligible, or observations of 0 were prevalent, then the Dirichlet-multinomial would be a reasonable choice. The degree to which the simulated data in Fisch et al. (2021) resembled the Georges Bank haddock data was not discussed. Also, Fisch et al. (2021) explored a different form of the logistic normal likelihood than is implemented in WHAM. It was noted that for the data in this assessment, the prevalence of zeros in the age composition was low (5 out of 792 observations in catch, 12 out of 387 in NEFSC spring, 49 out of 513 in NEFSC fall, and 2 out of 225 for DFO). Also, the zeros were not randomly distributed, rather they tended to be at older ages in years with low abundance. Furthermore, it was thought that process error was not likely to be small, and correlation was likely to be substantial. For this reason, the working group endorsed proceeding with the logistic-normal likelihood that treats zeros as missing.

From the wide array of models considered in this step, three contenders emerged:

- a 2 block model for fleet selectivity with no random effects in the first block (1931-1962, the years prior to a fishery-independent index being available), 2DAR1 random effects in the second block (1963-2019), logistic fleet selectivity in both blocks, and 2DAR1 random effects in NAA
- as above, but a single fleet selectivity block with 2DAR1 random effects in all years
- the single fleet block with 2DAR1 random effects in all years, and NEFSC spring and DFO indices with logistic selectivity

The working group decided that the 2 block model for fleet selectivity, with no random effects in the first block and 2DAR1 random effects in the second block, was a reasonable base model. The selectivity blocks explored in earlier ASAP and WHAM model configurations were specified to remove residual patterns but do not correspond with changes in regulated mesh size or minimum fish size, suggesting that apparent changes in selectivity are stochastic deviations rather than a reflection of regulatory periods. Rationale for no random effects in the first block included parsimony to not estimate random effects during the period when only fishery data are available (1931-1962), acceptable model diagnostics, and retrospective pattern. Hindcasting with mean absolute scaled error (MASE; Kell et al 2021) showed all 3 contenders performed well, with little difference in their predictive skill. However, it was recommended to follow-up on the issue of model non-convergence with the developer of WHAM (T. Miller).

WHAM exploration part IV – dodging boundary solutions, model exploration redux

A resolution to the problem of index selectivity parameters hitting the upper bound of 1 when selectivity at-age was estimated was provided by T. Miller. In the assessment model, it is the product of index-specific catchability and selectivity that premultiplies numbers at age to achieve predicted indices. Therefore, it is possible to fix q at a large value (well above any reasonably expected estimate), so that the selectivity at age parameters are estimated on a scale where the maximum value is <<1, thereby avoiding boundary solutions. After the model has been fitted, once can then re-scale selectivities by dividing each index selectivity ogive by it's maximum value (returning it to a scale where full selectivity equals 1), and catchability can be rescaled by multiplying the previously fixed value by the index-specific maximum value. Without this q-approach, catchability were typically on the order of 10^{-4}. When applying this approach, q was fixed at 2, 5, and 10 to see if results were impacted at all by the assumed fixed value – they were not. Alternatively, one can use the information about the age with maximum selectivity (from fixing q) and then re-fit the model by estimating q and fixing just the previously identified ages with maximum selectivity. A comparison of re-scaling q and selectivity versus re-fitting the model with the single age fixed produced identical solutions. For subsequent exploration, this "fixed-q" approach was implemented with all index catchabilities fixed to 2.0, and then selectivity and q are re-scaled after the model was fit.

A last and final round of model testing was conducted to ensure that previously discarded model structures (where a boundary solution prevented convergence) did not provide superior fits once the convergence issue was settled. These explorations encompassed:

- 6 levels of random effects in NAA (iid or AR1 in year for recruits only; or iid, AR1 in year or age, and 2DAR1 for all NAA)
- 4 levels of random effects in fleet selectivity (iid, AR1 in year or parameter, and 2DAR1)
- natural mortality was fixed at 0.2

From these 24 models, iid random effects in NAA or selectivity performed the worst, and this factor level was dropped prior to an additional model fitting iteration that explored natural mortality estimation:

- 5 levels of random effects in NAA (AR1 in year for recruits only; or AR1 in year or age, and 2DAR1 for all NAA)
- 4 levels of random effects in fleet selectivity (none, AR1 in year or parameter, and 2DAR1)
- 10 levels for natural mortality estimation (constant across all ages/years, or age specific across all ages/years) X random effects levels (none, iid, AR1 in year or age, and 2DAR1)

The results of this final exploration produced four models with fairly close AIC scores. In the context of AIC weights, there would be almost no probability assigned to a model differing by more than 4 from the best model. The four models summarized below differed by 1.8-6 AIC points; model nomenclature reflects model components that differed from the BASE (best) model:

1. BASE - The previously identified base model (which received working group consensus during exploration part III) again had the lowest AIC score. This model structure estimated NAA random effects as 2DAR1 (with a separate variance for recruits, and one for ages 2-9+), logistic fleet selectivity with 2DAR1 random effects in the second block, and natural mortality fixed at 0.2.
2. BASE_Mest - The next closest model ($\Delta AIC = 1.8$), had 2DAR1 random effects on NAA and fleet selectivity in the second block, and estimated a constant value for M (estimate was M=0.24)
3. BASE_AR1y_NAA - The third closes model ($\Delta AIC = 4.2$), had AR1 random effects across years in NAA and 2DAR1 random effects for fleet selectivity, and natural mortality was fixed at 0.2
4. BASE_Mest_AR1y_NAA fourth model ($\Delta AIC = 6$), had AR1 random effects across years in NAA and 2DAR1 random effects for fleet selectivity, and estimated a constant value for M (estimate was M=0.24)

Remaining models in this suite of runs differed by $\Delta AIC > 10$, failed to converge (all models estimating M at age), or had worse retrospective patterns. For models 2 and 4 described above, where M was estimated, there was much greater uncertainty in time series estimates of SSB, F, and recruitment compared to models 1 and 3 where M was fixed (Figure 122), although the retrospective pattern was halved (Table B20).

After all of the factorial explorations for model configuration, the base model structure defined in “exploration part III” consistently performed best. AIC was not the sole model selection criterion; model convergence and diagnostics, retrospective pattern, and hindcasting (only for a limited subset of model contenders) were also examined; jittering initial conditions was only explored for the proposed base model. In the case of winnowing the large factorial model structure explorations, non-converged models were dropped with no effort to look at diagnostics. Convergence was determined by tnlminb returning a convergence code of 0, and obtaining a positive-definite hessian. Maximum gradients for converged models were generally on the order of $1E-8$ - $1E-13$ Remaining converged models were sorted by AIC, and the magnitude of differences in AIC from the best model was examined simultaneously with values for Mohn’s rho. Often, trends in AIC and rho were associated with model structure for random effects (see **wham.explore.final.xlsx**). Models with the least retrospective pattern all had random effects in numbers at age for all age classes (naa_sig=“rec+1”) and among these the 2DAR1 correlation had the lowest AIC. Within models of comparable structure (random effect estimated or not), those with 2DAR1 random effects for fleet selectivity had vastly better AIC than those with ar1 or iid.

WHAM model diagnostics and results

Across the 4 models retained for detailed analysis (BASE, BASE_Mest, BASE_AR1y_NAA, BASE_Mest_AR1y_NAA), fits to the aggregate catch and indices were indistinguishable. Catch was matched more closely (Figure 123, and see similar plots named “Catch_4panel_fleet1.png” in each model folder plots_png), due to the tighter CV imposed on relative to the indices. The large peaks at the end of the index time series are fitted appropriately on log scale, even if on the observation scale one’s eye is drawn to seemingly large positive residuals (Figures 124, Figure 125, and Figure 126; and see plots named “Index_4panel_1.png”, “Index_4panel_2.png” and “Index_4panel_3.png” in each model folder plots_png). One-step ahead residuals are the appropriate diagnostic for state-space models (Berg and Nielsen 2016), and in WHAM these are currently produced for aggregate quantities (total catch and index mean in a given year) but are in development for age composition. The one-step ahead residuals are independent and should have no trend. There was no significant trend in the one-step ahead residuals for any of the 4 models (slopes not significantly different from zero; Figure 127 and see similar plots named *_osa.fleet.agg.plot.png in the root folder for each model).

Hindcasting is a form of cross validation, where observations are removed from the terminal year to evaluate prediction skill (Kell et al. 2021). Hindcasting was performed for all 4 models to evaluate ability of the models to predict missing index observations at the end of the time series. For each index (NEFSC spring, NEFSC fall, and DFO), a 4 year horizon (typically the period for short-term catch projections), the aggregate index and associated age-composition were removed for the last 1, 2, 3, or 4 years and the model was refitted. The mean absolute scaled error (MASE; Kell et al. 2021), was calculated and compared to the naive forecast that assumed the index remained at it’s previous value. The MASE is scale invariant and comparisons can therefore be made across data sets (Kell et al. 2021). A lower MASE score is better, and interpretation of a score of x is that it performs $(1/x)$ times better than the naive forecast – therefore, MASE scores <1 are superior to the naive forecast. As with other diagnostics, all 4 models performed very well, with a barely perceptible lower MASE score for the proposed base model (Figure 128). Looking across all time horizons and all 4 models, the lowest MASE scores were associated with the NEFSC spring index, and highest scores for the DFO index. This is due to the DFO survey only having an observation in 2 out of the 4 years of hindcasting.

The pattern of estimated random effects in abundance at age is very similar for these 4 models (Figure 129). The estimated random effects at age 1 show the early years of the model with slightly positive deviations, a period of negative deviations in the 1980s and 1990s, and both positive and negative deviations in the 2000s (with distinct large positive deviations associated with 2003, 2010, and 2013. These trends appear to propagate through ages 2-9 usually along the diagonal, suggesting cohort effects. The most recent five years have more negative deviations than the adjacent years.

Patterns in estimated selectivity for the fleet and indices was identical for these 4 models (Figure 130). There are two periods towards the end of the time series with reduced selectivity up to ages 4-5, which is likely responding to the reduced size due to density-dependent growth. The NEFSC spring survey has full selectivity at age 8, and slighly reduced selectivity at age 9 (~0.9); NEFSC fall survey has full selectivity at age 1 and declines to about 0.8 at the oldest ages; DFO survey has full selectivity at age 9.

Comparisons across the models in terms of estimated time series for average F, SSB, and recruitment, show very similar trajectories and scales (Figure 131). The main distinction among these runs are the greater confidence intervals associated with the two models that estimated M.

Model diagnostics and results discussed are very similar among the 4 models, yet the value for Mohn's rho was about halved for the models where M was estimated (Table B20). To understand what was driving this result, model estimates of parameters across the 7 retrospective peels were plotted (Figure 132; see files of type *_Retro_Est_Peels.pdf in the top level of each WHAM model directory). The estimates of fleet selectivity at age change as data are peeled from the terminal year, indicating different estimates of random effects on fleet selectivity for the shorter time series. Catchability also changes, but the pattern differed between models that estimated M or not; when M was estimated, retrospective catchability estimates got consistently larger, while for the model where M was constant, retrospective catchability estimates showed some increases and some decreases. A major source of change in the retrospective peels was the estimated value for natural mortality–in models that estimated it, retrospective estimates of M consistently got smaller, with values just about 0.2 in the model with data through 2019, down to <0.05 for the model with data through 2012. The changes in estimated deviations in abundance at age were indistinguishable (Figure 133), meaning the M changes were additive to rather than compensatory to the abundance deviations. Thus, by estimating natural mortality, the model is able to absorb additional structural inconsistencies between adjacent years; however, the values of M in those retrospective peels are unrealistic.

Although most other diagnostics seem comparable among the models, the examination of model estimates in the retrospective peels puts more credibility on the two models that keep M constant at 0.2 (WHAM_BASE and BASE_AR1y_NAA). Between those 2 models, WHAM_BASE had the lowest AIC (by 6.4) but BASE_AR1y_NAA had much better estimates of Mohn's rho; hindcasting scores favor WHAM_BASE but it is an almost imperceptible difference. As a final diagnostic, self-tests were conducted for these two models, where data were simulated from the fitted models with observation and process error added to the original fit. Relative errors were calculated for estimated quantities of total catch, SSB, average F, and recruitment, and calculated as $(sim.fit_i - sim_i)/sim_i$, where $sim.fit_i$ is the estimated quantity from the fitted model for simulated dataset i and sim_i is the true simulated value for dataset i. The BASE model was unbiased with 97% of models converging in this self-test, while the BASE_AR1y_NAA model showed median bias as high as 50% and only 70% of the models converged (Figure 134). There was lack of consensus among working group members as to the proper framework for self-testing state-space models. Extensive testing for the Georges Bank haddock assessment models generally showed that models that included 2DAR1 correlation structures performed well when both observation and process error were simulated but were biased for observation error-only simulations, while less complex correlation structures (*iid*, e.g.) performed well in both cases. This topic is recommended for future research.

The WHAM_BASE model is proposed as the preferred model. The 'peeled' models, which have from 1 to 7 years of data removed, were each re-fit and a 7-year retrospective analysis was done in order to characterize how the retrospective pattern has evolved in the last 7 years. For the model with 7 years removed, the terminal year is 2012. One or more of the peels in the retrospective analysis for this model did not converge and no estimate of Mohn's rho is available, however for the remaining peels, rho ranged from 0.12 to 0.31 for SSB and -0.1 to -0.35 for average F. The rho values were generally increasing from 2013 to 2018 with a slight decrease for 2019, similar to the pattern that was seen in the VPA but with greatly reduced magnitude (Table B21).

The terminal year estimate of SSB is 137.9 mt, with 95% confidence interval of [93.63, 203.10]. The rho-adjusted SSB is 105.7 mt, and falls within the 95% confidence interval. The terminal year estimate of average F is 0.146 with 95% confidence interval [0.0913, 0.232]. The rho-adjusted value for average F is 0.225, and falls within the 95% confidence interval. Time series estimates of SSB and F were compared for WHAM_BASE, ASAP_4Block_ADJCV_NEFF, ASAP_4Block_Flat_ADJCV_NEFF, and the final VPA model with all data adjustments (Base2019_Len-Cal.CAA.DFO.Y41) in Figure 135. Trends in F are very similar over the time series, and in most years, the ASAP and VPA estimates are within the 95% CI of the WHAM_BASE model. Trends and scale are also very similar for the time series of SSB, except for the very end of the time series, where the VPA had a value of 0.468 for Mohn's rho on SSB (which would scale the terminal SSB point from 667.9 mt to 455 mt).

WHAM sensitivities

Several individual sensitivities to the base model were conducted to test model structure and estimability of additional variance parameters or variance scalars:

- the base model estimates 2 variances and 2 correlation parameters for NAA random effects (one for age 1, and one for ages 2-9); three additional model runs were testesd: estimated of 3 sigmas (age 1, 2-3, and 4-9), and two versions estimating 4 sigmas (age 1, 2-3, 4-7, 8-9) and (age 1, 2-3, 4-5, and 6-9). When estimating additional sigmas, the estimate for age 1 was stable, but the estimate for sigma on ages 2-3 doubled while the sigmas estimated for ages 4 and older were about the same magnitude as the base model estimate for ages 2-9. The exploration was made to evaluate if it made any impact on the pattern and magnitude of estimated NAA random effects; it did not.

- a model run that attempted to estimate random effects on indices was done, but it produced imperceptible differences in the estimated index selectivity

Retrospective patterns: potential causes and plausability

Across the modeling platforms and structures explored, there was a consistent pattern in the trade-off between likelihood-based diagnostics and consistency diagnostics (retrospective pattern). More flexible models improve likelihood-based diagnostics and residual patterns, but at the cost of less precise estimates that are also less stable as years of data are peeled back; less flexible models generally have fewer parameters with greater precision and stability as data are peeled back, but likelihood diagnostics and residual patterns may not be so great. Potential factors contributing to the retrospective pattern are listed, along with an attempt to ascribe plausability.

- DFO survey does not cover deeper strata off of Georges Bank, where the haddock move seasonally, particularly in recent years; this creates an inconsistent index of the population, however the seasonal movements are more pronounced in fall (low?)
- changes in survey selectivity due to observed changes in growth were hypothesized, but models with RE in survey selectivity did not show any effect (low?)
- age smearing (as detected in the 2017 Georges Bank update assessment) into age classes adjacent to the exceptional year classes (low-medium, depending on year?)
- landings reporting (both over and under), unobserved discarding, misassignment of catch to statistical area/stock area (low-medium?)
- a change in M was hypothesized, but ageing data show some of the oldest fish in the population now, despite high abundance; but if M were changing recently that could produce a retro (unknown?)
- possible changes in availability of fish to the fleet; extensive exploration of fleet selectivity at age consistently suggested flat-topped selectivity, but any incidence of doming could create misspecification (unknown?)

Suggestions for future research

- the multinomial likelihood seems pre-disposed to find domes; this should be further tested
- structured simulations to evaluate ability of state space models to assign process error magnitude and cause
- further simulations to evaluate appropriate framework for self-tests
- a research track assessment on state-space models in WHAM is underway, and these and many other topics will be explored

TOR 5: Update or redefine status determination criteria (SDC point estimates or proxies for B_{MSY}, $B_{THRESHOLD}$, F_{MSY} and MSY) and provide estimates of their uncertainty. If analytic model-based estimates are unavailable, consider recommending alternative measurable proxies for BRPs.

The previous reference point approach for Georges Bank haddock dates back to the extensive work done at the Groundfish Assessment Review Meetings (NEFSC 2008). For Georges Bank haddock in particular, it was noted that density-dependent changes in size at age have implications for the assumption of stationarity of biological reference points, and that a number of selectivity blocks would be required if an SCAA rather than VPA formulation was considered. The VPA was selected as the preferred model in the last benchmark stock assessment, and reference points were based on long-term projections from a distribution of initial conditions (NEFSC 2008). Specifically, 1000 bootstrap replicate data sets were created by resampling index residuals and refitting the VPA model. From these fitted VPAs, there were 1000 estimated numbers at age (NAA) in the terminal year plus one (T+1, i.e., one year after the last catch at age data). From these NAA in T+1, the NOAA Fisheries Toolbox program AGEPRO was used to sample from a subset of recruitment estimates from the full time series of estimated recruitment in the base model. The subset was defined as recruitment estimates that are associated with spawning biomass above 75,000 mt, based on an odds ratio that suggested higher recruitment is observed when spawning biomass exceeds that threshold. Future recruitment was projected based on the empirical cumulative distribution function associated with this subset of recruitment estimates. Long term projections (100 years) were made with F=F40% as a proxy for FMSY, and using a 5-year average of selectivity, weights at age, and maturity. Results were inspected to verify convergence to approximate stability at year 100, and the medians are reported as proxies for SSBMSY, MSY, and RMSY. In 2008, there was debate as to whether the exceptional 1963 and 2003 year classes should be included, given that they appeared to be rare events (NEFSC 2008). Projections at the 2008 and 2012 assessment excluded them (NEFSC 2008, 2012), but assessments in 2015, 2017, and 2019 included them (NEFSC 2015, 2017, 2019).

The proposed base model is in WHAM, which has built-in projection capabilities. In calculating the reference points, WHAM can propagate parameter uncertainty into the FMSY and BMSY proxies. U.S. national standard guidelines specify that MSY reference points or proxies should reflect “prevailing ecological, environmental conditions and fishery technological characteristics (e.g., gear selectivity), and the distribution of catch among fleets” (NOAA 2016). The working group agreed to use the full time series of recruitment estimates and a recent 5 year average of selectivity, weights, and maturity at age. Use of the full time series of recruitment is justified by examining the trend in average recruitment over different time intervals, from a recent 5 year average to the full time series, in increments of 5 years. Short term average recruitment is strongly influenced by the recent exceptional year classes (Figure 136), while windows from the most recent 35 years through the full time series show a fairly stable mean recruitment and very little difference in the empirical cumulative distribution (cdf). For selectivity, examining the mean age-specific value over the same temporal windows as average recruitment, there is very little difference in selectivity at ages 1 and 6-9+ (Figure 137). Selectivity for ages 2-5 shows increasing values from longer time series. The shorter window seems more appropriate and reflects the impact of both changes in growth and changes in management (e.g., minimum size). For these reasons, a 5 year recent average appears justifiable. For similar reasons, a recent 5 year average is more reflective of population growth and would likely make a more reasonable basis for calculating YPR, SSB/R (Figures 138 and 139). The maturity ogive has 2 stanzas, and the final stanza is 4 years; for consistency, a 5 year average is recommended for maturity at age.

A study to examine performance of projections and catch advice for stocks assessed by the Northeast Fisheries Science Center was recommended by the New England Fisheries Management Council in 2011. Results of this study indicated that retrospective patterns were the leading cause of poor performance of projections, and that even after a rho-adjustment, the length of projections degraded after about 3 years (Brooks and Legault 2016). Consequently, management track assessment updates are performed on a 2-year cycle (for the most part), and reference points are routinely updated at the same time. This affords an opportunity to re-examine decisions on reference point specification, and any proposed changes would typically warrant a level-2 review (i.e., intermediate scope of peer review)..

TOR 6: Define the methodology for performing short-term projections of catch and biomass under alternative harvest scenarios, including the assumptions of fishery selectivity, weights at age, maturity, and recruitment.

Methodology for short term projections for Georges Bank haddock has been adjusted in nearly every management update since 2012 in order to deal with changes in growth (Figure 140) that impact projected weights at age and selectivity at age. The approach to make these adjustments was to compare projection assumptions made for weights and selectivity in the previous assessment with realized weights and selectivity for the current assessment. This led to reducing the window of recent average from 5 to 2 years, and treating the large year classes with year/age specific assumptions. Most recently, the weights at age for the 2013 year class in 2017 and 2018 (ages 4 and 5) were projected from the fit of a log linear model to weights at ages 1-3, and weights at ages 8 and 9 were assigned the time series minimum (NEFSC 2019). Selectivity was a 5 year average for most projection years/ages with several exceptions:

- Selectivity of the 2013 year class was assigned the same selectivity at age as the 2010 year class
- age 7 was assumed to be fully selected
- selectivity of ages 8, 9 was assumed equal to the average selectivity of ages 5-7 (consistent with VPA model formulation)

For this research track assessment, assumptions for weights at age were re-visited as follows:

- bias between assumed versus observed weights at age were compared for windows of 2, 3, 4, or 5 years
- bias between predicted versus observed weights for log linear model fits were compared for ages 4, 5, and 6 (based on fitting to ages 1, 2,3) and for ages 7, 8, 9 (based on fitting to ages 4, 5, and 6) for the 2003, 2010, and 2013 year classes
- to assess the combined effects of bias in predicted weights at age given numeric abundance at age, the different predicted weights at age were used to calculate total catch biomass and total spawning biomass as follows:
  - given model predicted NAA and F and M at age, total predicted catch was calculated as (NAA)x(FAA/ZAA)x(1-exp(-ZAA))x(predWAA) and compared to (NAA)x(FAA/ZAA)x(1-exp(-ZAA))x(obsWAA)
  - given model predicted NAA, total predicted spawning biomass at age was calculated as (NAA)x(maturityAA)x(predWAA) and compared to (NAA)x(maturityAA)x(obsWAA)

The results of these analyses revealed that the shorter the window, the less bias between assumed versus realized weights at age (Figures 141 and 142). This is not surprising given the declining weights at age trend (Figures 143 and 144). This calculation was repeated over a time span of 15, 20, or 30 years, and results were robust to the number of years considered, and also robust to whether or not the large year classes were included. The working group endorsed using a 2 year average for weights at age in the projection, but wanted to see comparisons between the log linear predicted weights that are currently used for the exceptional year classes.

Results of the log linear predicted weights at age show some skill at predicting ages immediately after those that were fitted, but bias in the predicted values increases at the older ages, particularly the plus group (Figures 145 and 146). The bias in this trend from the linear models likely explains the bias in calculating total catch biomass and total spawning biomass (Figure 147).

The final comparison evaluated the predicted weights for the various methods just for the 3 recent exceptional year classes. For the 2013 year class, the linear model fitted to ages 1, 2, and 3 performed well for predicting weights at ages 4 and 5 (as noted in the 2019 management update), however it did not perform as well on the 2003 or 2010 year class; in general, the 2 year average performed as well as or better than any of the log-linear model fits. The log linear model fitted to ages 4, 5, and 6 to predict ages 7, 8, and 9, did not perform well at all (Figure 148).

Overall conclusions from these analyses suggest that using a 2 year average of recent weights at age (up to age 8 at least) is the most robust assumption for short-term projections. With respect to assumed weights at age for the plus group, the ratio of observed weight at age 9 to observed weight at age 8 was examined (Figure 149). This trend varied between about 0.9 and 1.4 across all year classes in the model, but for the 2003 and 2010 year classes the average value was 1.06 and 1.05 for catch and ssb weights, respectively. Thus, the recommendation for projection weights at age 9 for exceptional year classes is to use the minimum of a 2 year average or the mean ratio of age 9 to age 8 (1.06 or 1.05); for other year classes, a 2 year average for plus group weight is expected to perform as well as any other approach.

With respect to assumed selectivity in projections, the model structure is vastly different from the VPA. Analyses about trends in selectivity by the window of years averaged over (see TOR5), suggest that the current 5 year average will capture relevant recent exploitation characteristics. As the selectivity is logistic, no special assumption is needed for the oldest ages.

Numbers at age in the base model have a 2DAR1 autoregressive structure on random effects, and the working group agreed to let the model propagate these into the future. The NAA deviations at the end of the model are not large, and revert to zero within 2-3 years (Figures 150 and 151). As noted above for ToR5 (reference points), management track assessment updates offer opportunities to re-examine projection specifications.

TOR 7: Review, evaluate and report on the status of the Stock Assessment Review Committee (SARC) and Working Group research recommendations listed in most recent SARC reviewed assessment and review panel reports. Identify new research recommendations.

2008 GARM III Assessment

The working group reviewed the following research recommendations from the 2008 GARM III GB haddock assessment, which was the last benchmark assessment for GB haddock (NEFSC 2008).

- It was observed that growth appears to be a function of density. As the data to examine this relationship is in the assessment, it should be investigated. Furthermore, if the effect is significant, it should be included in the BRP estimation.
  - This research recommendation was addressed under TOR 10 of this research track assessment, through an update of a growth analysis from GARM III.
- A good correlation was observed between chlorophyll and recruitment strength, especially the strong 2003 year - class. A similar correlation has been observed for other haddock stocks (e.g. Eastern Scotian Shelf haddock; Platt et. al, 2003). The Panel encouraged investigation of other potential covariates of the various aspects of production (growth, recruitment, and natural mortality).
  - This research recommendation was partially addressed under TOR 9 of this research track assessment, through an analysis of fall algal blooms and a drifter study. These efforts focused on covariates affecting recruitment. Additional work is needed to look at covariates affecting growth and natural mortality.

2019 Management Track Assessment

The working group reviewed the research recommendations from the 2019 GB haddock management track assessment, which was the most recent assessment for GB haddock (NEFSC in preparation).

Assessment Report

The following research recommendations are from the 2019 GB haddock management track assessment report (NEFSC in preparation).

- Projection advice and reference points for Georges Bank haddock are strongly dependent on recruitment. A decade ago, extremely large year classes were considered anomalies (e.g., 1963 and 2003). However, since 2003, there have been four more extremely large year classes (2010, 2013, 2016, and 2018). Future work could focus on recruitment forecasting and providing robust catch advice.
  - This research recommendation was partially addressed under TORs 5 and 6 of this research track assessment. While the WG did discuss how to handle recruitment when calculating reference points and conducting short term projections, there is more work that can be done to address this recommendation.

- Assumptions about weights at age and selectivity are very influential in short term projections. As multiple large year classes move through the population, it is difficult to predict how strong the density dependent response will be, but future work could continue examining performance of projected values with realized values.
    - This research recommendation was partially addressed under TOR 6 of this research track assessment. The WG reviewed and decided to use a log linear regression approach to calculate weight at age for strong year classes, but alternative analytical approaches should be explored.
- For this assessment, reference points are estimated with a recent 5 year average for selectivity, maturity, and weights at age, whereas short-term projections use year-specific decisions to deal with the current large year classes. Considering that estimated population abundance at MSY is much less than the current population abundance, recent average biological and selectivity parameters may not reflect MSY conditions. Calculating per recruit statistics on an annual basis demonstrates the dynamic range of reference points in response to density dependent changes in growth.
    - This research recommendation was addressed under TORs 5 and 6 of this research track assessment. The WG evaluated a range of years to use when averaging selectivity, maturity, and weights at age for reference points and short term projections.

Review Panel Report

The following research recommendations are from the 2019 GB haddock management track review panel report (NEFSC in preparation).

- A subset of this stock is also assessed by the Transboundary Resources Assessment Committee (TRAC) for the eastern portion of the stock only. Both assessments assume a closed population, which cannot be true for both. Previous research on stock identification and the current resource distribution suggests that the entire Bank should be considered a unit stock.
    - This research recommendation was partially addressed under TOR 12 of this research track assessment. The WG reviewed previous studies that looked at GB haddock stock structure, but additional work (e.g., genetic studies) could be done on this topic.
- During the upcoming research track in 2021, statistical catch-at-age or state-space modeling approaches should be considered to allow improved tracking of survey indices and allow for uncertainty in catch at age (particularly for dominant year classes) and more control over fishery and survey selectivity estimation.
    - This research recommendation was addressed under TOR 4 of this research track assessment. The WG reviewed both statistical catch-at-age and state-space models for the GB haddock assessment.

Haddock Research Track Recommendations

- The WG recommends that the gutted:whole weight conversion factors for haddock be updated. The current conversion factors were developed in the 1930s. It is not certain how the factors were estimated and what data were used. In addition, it should be determined whether or not there is a difference in conversion factors for fish gutted by hand versus fish gutted by machine.

- For jointly managed US-Canadian stocks, survey strata sometimes cross stock boundaries. Since this issue affects multiple stocks, the WG recommends that a research track or TRAC workshop be held to evaluate different methods for handling these cross-stock strata.
- When an alternative assessment approach must be used, the WG recommends that analysts be able to explore analytical approaches as part of a management track assessment, because an alternative assessment approach should only be a temporary measure until a new analytical approach can be developed.
- The WG recommends that methods for estimating weight at age and fishery selectivity for strong year classes in short term projections be further explored. Density-dependent growth makes it difficult to forecast these quantities.
- There are differences between how the US and Canada collect the weight data used to estimate weight at age. For example, the US does not collect individual weights from the commercial fishery, while Canada does. Due to these differences, the WG recommends that work continue to evaluate different methods for calculating weight at age and incorporating weight at age data into assessment models.
- The US and Canada use different maturity stage classification systems. The WG recommends that the differences in classification systems be evaluated to determine what impacts these differences may have on the resulting maturity ogives. In addition, WG recommends that maturity data continue to be collected in the field.
- The WG's review of existing haddock stock structure studies proved inconclusive. The WG recommends a genetic study to compare haddock from different areas (e.g., GOM, GB, EGB, and 4X) to help inform haddock stock structure in the region.
- The WG found no new evidence to support a change in the assumed natural mortality rate of 0.2. The WG recommends that work be done to identify factors influencing the natural mortality of haddock on GB.
- The WG recommends that work be done to incorporate density-dependent processes: growth, spatial distribution and natural mortality into stock assessment models.
- The WG recommends that the method developed to avoid boundary constraints for selectivity parameters in these applications be implemented in the WHAM program for other applications.
- Further investigation of genetic variability of haddock from eastern Georges Bank and the Great South Channel would help to determine stock structure.

TOR 8: Develop a "Plan B" for use if the accepted assessment model fails in the future.

In the event that the proposed base model (WHAM_BASE) or alternative model configurations are not accepted as a basis for status determination and fishery management advice, the proposed alternative assessment framework is to use the Plan B Smooth. This has been the proposed back-up method since the 2017 management track assessment (NEFSC 2017), but has never needed to be applied. A research track working group investigating performance of index-based methods (NEFSC 2020 *in prep*; Legault et al., *submitted*), conducted a large-scale simulation study to understand if simpler methods would perform better in situations where an age-based assessment was rejected due to a large retrospective pattern (Mohn's rho of about 0.5 for SSB). Plan B smooth was one of the methods, and across all scenarios considered, its performance was generally robust, except for scenarios where retrospective patterns were caused by unreported catch and the stock was depleted. The Georges Bank haddock stock is not depleted. The index-based methods working group also applied a statistical catch at age model with an adjustment for retrospective pattern to the same simulated data, and found that it performed as well as the more robust index based methods and had the advantage of being able to determine stock status. Although a wide-range of age-structured models were explored in the development of the current base model, and could potentially serve as an alternative model, the exact model specifications of any proposed age-structured alternative would depend on the reason for rejection of the proposed base model. Lacking omniscience on a possible reason for rejection, the Plan B smooth is proposed, and justification for details of its implementation are summarized.

Briefly, Plan B smooth fits a loess (second degree polynomial) to average survey biomass, and then a log-linear regression is fit to the last 3 years of the loess fit. The slope from the log-linear regression is then exponentiated and becomes a multiplier on either catch or quota. The main decisions to be made when applying the method are: i) indices to use; ii) loess details (years to use, 'family' option for the loess smoother, and the span which controls the degree of smoothing); iii) whether to impose a cap on the estimated multiplier (i.e., specify that the multiplier can't be larger/smaller than X%); iv) whether to apply the multiplier to catch or quota. The 'family' option in the loess function is to fit by least squares (family=Gaussian, assumes normal error) or to use a robust regression (family=symmetric) which performs additional fit iterations that downweights outliers. There is interaction between the number of years and the span supplied to the loess function, and it is important to evaluate consequences of those decisions. The span parameter in the loess is a proportion between 0 and 1 and controls the degree of smoothing, with larger values producing smoother lines.

On the first decision, there are three indices available to use in the method: NMFS-Spring, NMFS-Fall, and DFO. The DFO survey did not sample all strata on Georges Bank in 1993, 1994, 2005, 2006, 2010, 2012, 2015, 2017, 2018, and 2020, and the working group agreed to just use the two NMFS surveys. With respect to the second decision, the group decided to not include the years where the Yankee 41 net was used for the NMFS Spring survey. This leaves a complete time series for spring and fall from 1982-2019 (a total of 38 years that could be fitted). Loess fits to the average time series were evaluated across multiple spans and numbers of years (Figure 152). Given the strong swings in abundance as large year classes transit through the population, there were clear limitations to large span values, which were not able to match the trends in the index. Examining initial loess fits allowed a closer examination of spans (0.21-0.36 in increments of 0.03) and number of years (n=33 or 38) that appeared to capture the essential dynamics. Final selection was made by looking at the retrospective pattern of loess fits as 7 years of data were sequentially removed from the terminal year of the average index, and corresponding retro pattern in the estimated multiplier (Figures 153 and Figure 154). The working group decided to use a window of 33 years of data and a span of 0.27. This means that in future applications, the window of 33 years would slide forward and the Plan B method would be fit to the most recent 33 years of data.

The last two working group decisions have to do with application of the estimated multiplier. Catches have been far below quotas for this stock, and applying the multiplier to catch would implicitly penalize fleets for this. The working group agreed that the multiplier should be applied to quota. However, it was noted that existing quotas likely have a scale issue (most likely overestimated), given the combined effects of the new survey calibration and changes in the Canadian catch at age and weights at age. The working group discussed imposing a cap on the multiplier, as a way to avoid wild swings, which are expected when the loess is faced with smoothing over a large peak and subsequent decline. No direct decision was made on a multiplier cap, but discussion of future work to evaluate the performance of multipliers was recommended.

TOR 9: Review and present any research related to recruitment processes (e.g., spawning and larval transport, and retention), and potential hypotheses for large recruitment events.

Fall bloom

The Working Group considered an update to the analysis of the effect of the fall bloom on haddock (Melanogrammus aeglefinus) recruitment. It has been hypothesized that the recruitment of haddock on Georges Bank (GB) may be influenced by the provisioning effects of the fall bloom on pre-spawning adults (Friedland et al. 2015). Since haddock tend to feed on detritivores like ophiuroids and amphipods, the mass flux of fixed carbon associated with a phytoplankton bloom may affect the energy flow to haddock on relative short time scales. The fall bloom would provide energy that haddock could invest into reproduction affecting either or both fecundity and egg condition. Haddock females are also known to skip spawning in years with limited energy resources, which adds a further dimension to how provisioning may affect recruitment (Skjæraasen et al. 2020).

This hypothesis has stimulated considerable debate and productive discussion (Friedland et al. 2009). The original analysis supporting this hypothesis was based on a sample size of n=7 and ran contrary to more classic hypotheses of recruitment control related to first feeding and larval retention. As is the case in fisheries and related fields, correlation may not always detect robust mechanistic linkages (Havens 1999), so there is a responsibility to monitor and test correlative relationships as new data becomes available. This is particularly important when there has been a change in the conditions or events that may affect the correlation. This may be the case with the GB haddock stock, which experienced the highest recruitment on record in 2013, a year class not included in previous analyses (Finley et al. 2019).

The updated analysis tested the patency of the relationship between GB haddock recruitment and the dimensions of phytoplankton blooms forming on GB in the fall prior to spawning (Friedland 2021). The analysis was successful in modeling the exceptional recruitment in 2013 and can thus be attributed to the provisioning hypothesis and more generally showed that relationship holds with a more than a doubling of recruitment data considered in the original analysis (Figure 155). The figure shows the relationship between the recruits per spawner ratio based on survey indices for the GB stock versus fall bloom magnitude the year prior to spawning. Bloom magnitude is the sum of chlorophyll concentration during the bloom period and is meant to represent the overall potential of mass flux to the benthos. The bloom detection, performed using a change point algorithm, was conducted used a single set of detection parameters in panel (a) of the figure and a range of parameters to increase detection rates in panel (b). The better fit was achieved using the range of parameters approach.

The Working Group concluded that the provisioning hypothesis may be contributing to the observed pattern of recruitment of GB haddock. Although there were three exceptionally strong year classes in the last two decades, only two of them (2003 and 2013) were identified to have high bloom magnitude. The bloom algorithm did not find a clean beginning and end to a bloom event associated with the 2010 year class.

Egg retention

The WG also considered ongoing research by Sheremet et al. (*In prep.*) on the retention of haddock eggs and larvae based on both real and simulated ocean drifter trajectories. Circulation on Georges Bank is clockwise, and the timescale of eggs and larvae development (~2 months) coincides with the timescale of flow circulation. Lough and O'Brien (2012) concluded that abundance of eggs retained depends on wind-driven transport. Sheremet et al. (*In Prep.*) used the Finite Volume Coastal Ocean Model (FVCOM; Chen et al. 2006), a prognostic, unstructured-grid ocean circulation model that assimilates hydrography, altimetry, atmospheric conditions, and tides, with hourly output, to study patterns and variability in retention. The results presented to the WG focused on understanding where eggs end up when spawned, and what factors influence larval drift.

For a weekly time step, 10,000 theoretical drifters, which serve as a proxy for spawned eggs, were released from the Northeast peak of Georges Bank and their trajectory was projected based on depth-averaged flows. Haddock eggs are neutrally bouyant and well mixed over the water column (Lough et al. 1996), so it is appropriate to use depth-averaged rather than surface flows. An initial comparison of retention after 62 days was made based on the climatological monthly-mean flow (averaged over 1978-2016), as well as for a yearly mean flow (averaged over all months and all years) that would imply that spawning occurred throughout the whole year. For drifters released on the NE peak of Georges Bank, the yearly climatology showed no retention while the monthly climatology showed that spawning in April would produce the highest retention (Figure B9.156).

Year-specific retention was calculated next, and it was found that retention fraction of the 2003, 2010, 2013 year classes was much greater than the mean retention fraction across all years (Figure 157). There are other years with high retention fraction (1995, e.g.), however no strong recruitment materialized. Several of these other calculated high retention years occurred in years immediately after the stock had been declared collapsed; it may be that despite conditions being ripe for high recruit retention, there were insufficient spawners in the population. It was concluded that high retention in general does not guarantee strong recruitment; spawning biomass is necessary to produce eggs, and after the egg/larval stage, many other factors can act post settlement. However, it is certainly true that high recruitment cannot occur with low retention.

These same analyses conducted for the Northeast Peak were repeated for releases in the Great South Channel, known to be another spawning location for haddock (Figure 158). Closed area I was used to define the Great South Channel, and the analyses performed for the NE peak were repeated. As with releases on the Northeast Peak, retention was greatest for release in April. Most drifters released in April are retained in the Georges Bank gyre, and therefore, connectivity at egg/larval stage is expected across all of GB. All drifters that occasionally digress into Gulf of Maine eventually return to Georges Bank. Given observed current flows, the essential boundary in the region is in the middle of the Great South Channel across which the drifters permanently leave the area. Retention of the 2003, 2010, and 2013 year classes was higher than the average, but the difference was not as dramatic as for releases on the Northeast Peak.

Lastly, an examination of real drifter tracks revealed the following patterns:

- No drifters that go through the NE Peak make it back into the Gulf of Maine
- No drifters that go through the NE Peak make it back into the shoals east of Nantucket
- Nantucket Shoals spawn might often reach NE Peak
- Drifters emanating from the GSC generally go to Georges Bank, only a few exit the bank to the north or to the Mid-Atlantic Bight
- Drifters emanating from the North Shore generally go to Georges Bank, few make it north and east of the NE peak, and most eventually make it to the Gulf Stream
- Drifters from the Western Gulf of Maine generally stayed in the Gulf of Maine, but a few made it to Georges Bank and a few travel to the north

These results align with the simulated drifter results, and other information reviewed in TOR12. There are several caveats when comparing the real drifters and simulated drifters. The real drifters are opportunistic releases in space and time, not a designed study; surface velocities were seen to be ~40% faster than depth-averaged in FVCOM simulations so distance traveled may be greater than expected; and the number of drifter observations during haddock spawning time is limited.

TOR 10: Review and present any research related to density-dependent growth.

No directed work to formally model density-dependent growth on Georges Bank was conducted, however discussion can be found in earlier TORs for dealing with the observed trends in size and weight, and how best to deal with those trends in reference points and short-term projections.

TOR 12: Review data related to stock structure of haddock on Georges Bank (including Eastern Georges Bank management area) and implications for assessments conducted on the whole bank and on subareas of the bank.

The Working Group's approach to this term of reference was to determine the most plausible biological stock structure for Georges Bank haddock that supports stock assessment model assumptions (e.g., reproductive isolation, demographic independence, homogeneous vital rates, no immigration/emigration) and to consider practical aspects of spatial stock assessment and management units. This term of reference does not include consideration of Gulf of Maine haddock stock structure.

The haddock fishery in the Georges Bank region is extremely data-rich and haddock stock structure has been studied in the region for nearly a century (e.g., previous reviews by Needler 1930, Grosslein 1962, Clark et al. 1982, Begg 1998, NEFSC 2014). However, stock identity remains uncertain, because stock boundaries are indistinct and geographic stock structure appears to be largely influenced by haddock abundance (Cadrin et al. 2020 Working Paper). The US and Canada initially managed haddock fisheries in the region as two spatial units (Western Scotian Shelf, and Georges Bank-Gulf of Maine; Figure 159; Bowen 1987), but since 1977 the US has managed haddock fisheries in the Gulf of Maine (areas 511-515) separately from those on Georges Bank and off southern New England (areas >515). A transboundary management unit on eastern Georges Bank (areas 551, 552, 561, 562) was established in 2001 (TMGC 2002) that is nested within the US Georges Bank management unit.

Large channels form partial barriers to movement of juvenile and adult haddock (Figure 160), and larval retention gyres also limit mixing of eggs and larvae across these channels (Figure 161). The oceanographic gyre on Georges Bank disperses eggs and larvae across the Bank, retaining most of them on the Bank, and occasionally dispersing others off the Bank (Walford 1938, Colton and Temple 1961, Smith and Morse 1985, Polacheck et al. 1992, Lough et al. 2006, Boucher et al. 2013). Recent oceanographic modeling and drifter studies confirm the dispersal of eggs and larvae across Georges Bank, with some dispersal from spawning on the northeast peak of Georges Bank to Southern New England but little connectivity at early life stages between Georges Bank and the Gulf of Maine (Sheremet et al. 2021 Working Paper). Transport rate of eggs and larvae is driven by physical oceanography, but effective transport among areas appears to increase during and following strong recruitment events (Grosslein and Hennemuth 1973, Overholtz 1985, Brodziak et al. 2008). The analyses developed by Overholtz (1985) for the 1975 and 1976 year-classes were applied to the recently dominant 2003, 2010 and 2013 year-classes and showed similar widespread dispersal of strong year-classes (Figure 162).

During periods of low abundance, the spatial distribution of juvenile and adult haddock was more discontinuous, with discrete concentrations on eastern Georges Bank, in the Great South Channel, in the Gulf of Maine and on the Scotian Shelf (Begg et al. 1999, Cargnelli et al. 1999, Brodziak 2005). However, during periods of high abundance, spatial distributions were more continuous between eastern Georges Bank and the Great South Channel (Figure 163, Sosebee and Cadrin 2006). Therefore, static boundaries that meet the unit stock assumption are difficult to define.

Geographic variation in genetic, phenotypic and demographic traits suggests an isolation-by-distance pattern, with occasional mixing of haddock from multiple areas (Zwanenburg et al. 1992, Berg et al. 2020). The US Georges Bank Management unit includes multiple discrete spawning components of haddock on eastern Georges Bank and the adjacent Great South Channel (Purcell et al. 1996, Lage et al. 2001). During a period of low abundance in the late 1980s-early 1990s, juvenile and adult haddock were discretely distributed on eastern Georges Bank and had faster growth and maturity rates than haddock in the Great South Channel (Begg et al. 1999).

In summary, there are no persistent stock boundaries of haddock on Georges Bank because the resource distribution and connectivity among areas are dynamic. Therefore, previous boundaries defined during periods of low abundance (e.g., Gavaris 1989) may need to be reconsidered. Length and age compositions of haddock were similar among eastern Georges Bank, western Georges Bank, and the Great South Channel, suggesting a well-mixed stock. Length and age distributions across statistical areas show strong correspondence in the mode, suggesting homogeneous mixing across the bank (Figures 37 and 40 in TOR2, and 93 - 96 in ToR3). However, exploratory stock assessment modeling did not indicate evidence of emigration from eastern Georges Bank (see ToR4).

Separate assessment of eastern Georges Bank haddock includes a relatively homogeneous resource but does not account for larger-scale recruitment dynamics, and emigration of juveniles and adults to western Georges Bank may be a factor in recent retrospective patterns. The current distribution and connectivity of haddock across the Bank suggest that haddock on Georges Bank (eastern Georges Bank and western Georges Bank) is a single stock. Despite some evidence of geographic variation and partial isolation, haddock in the Great South Channel cannot be considered a separate stock. From a practical perspective, there have been few landings of haddock on Nantucket Shoals and other areas off southern New England and in the Mid Atlantic Bight since the mid-1900s (e.g., ~2% of total US catch since 1964; Figure 164, See ToR2), so haddock in those areas are not an important component of the current resource and fishery.

References

Azarovitz, T. R. 1981. A brief historical review of the Woods Hole Laboratory trawl survey time series. In Doubleday, W.G. and D. Rivard, (eds.), Bottom trawl surveys. Can. Spec. Publ. Fish. Aquat. Sci. 58:62-67.

Begg GA. 1998. A Review of Stock Identification of Haddock, Melanogrammus aeglefinus, in the Northwest Atlantic Ocean. Marine Fisheries Review 60(4):1-15.

Begg GA, Hare JA, Sheehan DD. 1999. The role of life history parameters as indicators of stock structure. Fisheries Research 43:141-163

Berg PR, Jorde PE, Glover KA, Dahle G, Taggart JB, Korsbrekke K, Dingsør GE, Skjæraasen JE, Wright PJ, Cadrin SX, Knutsen H, Westgaard J-I. 2020. Genetic structuring in Atlantic haddock contrasts with current management regimes. ICES Journal of Marine Science 78: 1-13.

Boucher JM, Chen C, Sun Y, Beardsley RC. 2013. Effects of interannual environmental variability on the transport-retention dynamics in haddock Melanogrammus aeglefinus larvae on Georges Bank. Mar. Ecol. Prog. Ser. 487:201–215.

Bowen WD. 1987. A review of stock structure in the Gulf of Maine area: a workshop report. CAFSAC Res. Doc. 87/21. 57 pp.

Brodziak JKT. 2005. Haddock, Melanogrammus aeglefinus, life history and habitat characteristics, Second Edition. NOAA Tech. Mem. NMFS-NE-196, 64 pp.

Brodziak J, Traver ML, Col LA. 2008. The nascent recovery of the Georges Bank haddock stock. Fish. Res. 94:123-132. Brooks et al. 2010.

Brooks EN, Miller TJ, Legault CM, O'Brien L, Clark KJ, Garvaris S, Eeckhaute LV. 2010. Determining Length-based Calibration Factors for Cod, Haddock and Yellowtail Flounder. Transboundary Resource Assessment Committee (TRAC) Reference Document 2010/08. 23 p.

Brown, B.E. 1965. Conversion Factors Used by the United States in Submitting Statistics to I. C. N. A. F. Bureau of Commercial Fisheries Biological Laboratory Woods Hole. Massachusetts Laboratory Reference No. 65-8. https://apps-nefsc.fisheries.noaa.gov/rcb/publications/series/whlrd/whlrd6508.pdf

Cadrin S, Wang Y, Finley M, Brooks EN. 2020. A review of haddock stock structure in the Georges Bank region. Haddock Research Track Assessment Working Group, Working Document. November 25, 2021.

Conser, R. J. and J.E. Powers. 1989. Extensions of the ADAPT VPA tuning method designed to facilitate assessment work on tuna and swordfish stocks. ICCAT Working Doc. scrs/89/43. 15pp.

Chen, C, R. C. Beardsley and G. Cowles, 2006. An unstructured grid, finite-volume coastal ocean model (FVCOM) system. Special Issue entitled "Advance in Computational Oceanography", Oceanography, vol. 19, No. 1, 78-89.

Cargnelli LM, Griesbach SJ, Berrien PL, Morse WW, Johnson DL. 1999. Essential fish habitat source document: Haddock, Melanogrammus aeglefinus, life history and habitat characteristics. NOAA Tech. Mem. NMFS-NE-128.

Clark SH, Overholtz WJ, Hennemuth RC. 1982. Review and assessment of the Georges Bank and Gulf of Maine haddock fishery. J. Northwest Atl. Fish. Sci. 3:1–27.

Colton JB Jr, Temple RF. 1961. The enigma of Georges Bank spawning. Limnology and Oceanography 6:280-291. DFO (Canada Department of Fisheries and Oceans) 1998. Eastern Georges Bank Haddock. DFO Sci. Stock Status Rep. A3-08.

Fisch, N., Camp, E., Shertzer, K., Ahrens, R., 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fisheries Research 243, 106069. https://doi.org/10.1016/j.fishres.2021.106069

Francis R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68(6): 1124–1138.

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

Finley, M. et al. 2019. Transboundary Resources Assessment Committee: Assessment of Haddock on Eastern Georges Bank for 2019. in press.

Friedland, K. D. 2021. A test of the provisioning hypothesis of recruitment control in Georges Bank haddock. - Can. J. Fish. Aquat. Sci. 78: 655–658.

Friedland, K. D. et al. 2009. Reply to the comment by Payne et al. on "Does the fall phytoplankton bloom control recruitment of Georges Bank haddock, Melanogrammus aeglefinus, through parental condition?" - Can J Fish Aquat Sci 66: 873–877.

Friedland, K. D. et al. 2015. Layered effects of parental condition and larval survival on the recruitment of neighboring haddock stocks. - Can J Fish Aquat Sci 72: 1672–1681.

Friedland KD, Langan JA, Large SI, Selden RL, Link JS, Watson RA, Collie JS. 2020. Changes in higher trophic level productivity, diversity and niche space in a rapidly warming continental shelf ecosystem. Science of the Total Environment 704: 135270.

Gavaris S. 1989. Assessment of Eastern Georges Bank Haddock. CAFSAC Research Document 89/49.

Grosslein MD. 1962. Haddock stocks in the ICNAF convention area. ICNAF Redbook III:124–131.

Grosslein MD, Hennemuth RC. 1973. Spawning stock and other factors related to recruitment of haddock on Georges Bank. ICES Rapp. Proc. Verb. 164:77-88.

Havens, K. E. 1999. Correlation is not causation: a case study of fisheries, trophic state and acidity in Florida (USA) lakes. - Environmental Pollution 106: 1–4.

Hennemuth RC. 1969. Status of the Georges Bank haddock fishery. ICNAF Res. Doc. 69/90. 21 pp. Jensen AC. 1967. A brief history of the New England offshore fisheries. US Bureau of Commercial Fisheries, Fishery Leaflet 594.

Kulka DW. 2012. History and Description of the International Commission for the Northwest Atlantic Fisheries (https://www.nafo.int/Portals/0/PDFs/icnaf/ICNAF_history-kulka.pdf).

Lage C, Purcell M, Fogarty M, Kornfield I. 2001. Microsatellite evaluation of haddock (Melanogrammus aeglefinus) stocks in the northwest Atlantic Ocean. Can. J. Fish. Aquat. Sci. 58:982–990.

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

Le Cren, E. D. 1951. The length-weight relationship and seasonal cycle in gonad weight and condition in the perch (Perca flavescens). Journal of Animal Ecology. 20:201–219.

Lough RG, Hannah CG, Berrien P, Brickman D, Loder JW, Quinlan JA. 2006. Spawning pattern variability and its effect on retention, larval growth and recruitment in Georges Bank cod and haddock. Mar. Ecol. Prog. Ser. 310:193–212.

Lough, R.Gregory, Lough, Elaine M. Caldarone, Teresa K. Rotunno, Elisabeth A. Broughton, Bruce R. Burns, Lawrence J. Buckley, Vertical distribution of cod and haddock eggs and larvae, feeding and condition in stratified and mixed waters on southern Georges Bank, May 1992, Deep Sea Research Part II: Topical Studies in Oceanography, Volume 43, Issues 7–8, 1996, Pages 1875-1904, ISSN 0967-0645, https://doi.org/10.1016/S0967-0645(96)00053-7.

MacCall AD. 1990. Dynamic geography of marine fish populations. Washington Sea Grant Program Seattle, WA. 153pp. Pershing AJ, Alexander MA, Brady, DC, Brickman D., Curchitser EN, Diamond AW, Mclenachan L, Mills KE, Nichols OC, Pendleton DE, Record NR, Scott JD, Staudinger MD, Wang Y. 2021. Climate impacts on the Gulf of Maine ecosystem : A review of observed and expected changes in 2050 from rising temperatures. Elem Sci Anth. 9: 1–18.

Miller TJ, Das C, Politis PJ, Miller AS, Lucey SM, Legault CM, Brown RW, Rago PJ. 2010. Estimation of Albatross IV to Henry B. Bigelow calibration factors. Northeast Fish Sci Cent Ref Doc. 10-05; 233 p.

Needler AWH. 1930. The migrations of haddock and the interrelationships of haddock populations in North American waters. Contributions to Canadian Biology and Fisheries 6(1): 241-313

NEFMC (New England Fishery Management Council). 1985. Fishery Management Plan, Environmental Impact Statements, Regulatory Impact Review, and Initial Regulatory Flexibility Analysis for the Northeast Multi-Species Fishery. (https://s3.amazonaws.com/nefmc.org/MultiSpecies-FMP.pdf)

NEFMC (New England Fishery Management Council). 2009. Amendment 16 to the Northeast Multispecies Fishery Management Plan. (https://s3.amazonaws.com/nefmc.org/091016_Final_Amendment_16.pdf)

NEFSC (Northeast Fisheries Science Center). 1992. Report of the Thirteenth Northeast Regional Stock Assessment Workshop (13th SAW). NEFSC Ref. Doc. 92-02.

NEFSC (Northeast Fisheries Science Center). 1995. Report of the 20th Northeast Regional Stock Assessment Workshop (20th SAW). NEFSC Ref. Doc. 95-18.

NEFSC (Northeast Fisheries Science Center). 1997. Report of the 24th Northeast Regional Stock Assessment Workshop (24th SAW). NEFSC Ref. Doc. 97-11.

NEFSC (Northeast Fisheries Science Center). 1998. Report of the 27th Northeast Regional Stock Assessment Workshop (27th SAW). NEFSC Ref. Doc. 98-14.

NEFSC (Northeast Fisheries Science Center). 2002. Assessment of 20 Northeast groundfish stocks through 2001: a report of the Groundfish Assessment Review Meeting (GARM), Northeast Fisheries Science Center, Woods Hole, Massachusetts, October 8-11, 2002. NEFSC Ref. Doc. 02-16.

NEFSC Vessel Calibration Working Group. 2007. Proposed vessel calibration studies for NOAA Ship Henry B. Bigelow. NEFSC Ref. Doc. 07-12.

NEFSC (Northeast Fisheries Science Center). 2008. Assessment of 19 Northeast Groundfish Stocks through 2007: Report of the 3rd Groundfish Assessment Review Meeting (GARM III), Northeast Fisheries Science Center, Woods Hole, Massachusetts, August 4-8, 2008. NEFSC Ref. Doc. 08-15.

NEFSC (Northeast Fisheries Science Center). 2014. 59th Northeast Regional Stock Assessment Workshop (59th SAW) Assessment Report. NEFSC Ref. Doc. 14-09. 782 pp.NEFSC (Northeast Fisheries Science Center). 2019. Stock Assessment Update of 14 Northeast Groundfish Stocks Through 2018. NEFSC Ref. Doc. xx-xx (https://s3.amazonaws.com/nefmc.org/8_Prepublication-NE-GrndfshOp-Assessment-1-7-2020-revision.pdf).

NEFSC (Northeast Fisheries Science Center). In preparation. Operational Assessment of 14 Northeast Groundfish Stocks, Updated Through 2018. US Dept Commer, Northeast Fish Sci Cent Ref Doc. Overholtz WJ. 1985. Seasonal and age-specific distribution of the 1975 and 1978 year-classes of haddock on Georges Bank. NAFO Sci. Council Studies 8:77-82.

Palmer, M. 2008. A Method to Apportion Landings with Unknown Area, Month and Unspecified Market Categories Among Landings with Similar Region and Fleet Characteristics. Groundfish Assessment Review Meeting (GARM III-Biological Reference Points Meeting). Working Paper 4.4: 9 p.

Pershing AJ, Alexander MA, Brady, DC, Brickman D., Curchitser EN, Diamond AW, Mclenachan L, Mills KE, Nichols OC, Pendleton DE, Record NR, Scott JD, Staudinger MD, Wang Y. 2021. Climate impacts on the Gulf of Maine ecosystem : A review of observed and expected changes in 2050 from rising temperatures. Elem Sci Anth. 9: 1–18.

NEFSC Vessel Calibration Working Group. 2007. Proposed vessel calibration studies for NOAA Ship Henry B. Bigelow. NEFSC Ref. Doc. 07-12.

NEFSC (Northeast Fisheries Science Center). 2020 In preparation. Operational Assessment of 14 Northeast Groundfish Stocks, Updated Through 2018. US Dept Commer, Northeast Fish Sci Cent Ref Doc.

NOAA. 2016. Magnuson-Stevens Act Provisions, National Standard Guidelines. Federal Register 81(201): 71858-71904.

Pennington, M., and Vølstad, J.H. 1994. Assessing the effect of intrahaul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, 50(3): 725–732. doi:10.2307/2532786.

Sheremet, V. , Manning, J., Brooks, E.N., Lough, R.G. (*In preparation*). Retention-Recruitment Characterization of Fish Larvae on Georges Bank from a 39 Year Circulation Model Hindcast.

Skjæraasen, J. E. et al. 2020. Large annual variation in the amount of skipped spawning for female Northeast Arctic haddock Melanogrammus aeglefinus. - Fisheries Research 230: 105670.

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

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

Wigley, S.E., H.M. McBride, and N.J. McHugh. 2003. Length-weight relationships for 74 fish species collected during NEFSC Research Vessel bottom trawl surveys, 1992-1999. NOAA Tech. Mem. NMFS-NE-171.

Wigley S.E., P. Hersey, and J.E. Palmer. 2008a. A Description of the Allocation Procedure Applied to the 1994 to 2007 Commercial Landings Data. US Dept. Commer., Northeast Fish. Sci. Cent. Ref. Doc. 08-18: 61 p.

Wigley, S.E., M.C. Palmer, J. Blaylock, and P.J. Rago. 2008b. A Brief Description of the Discard Estimation for the National By-Catch Report. Northeast Fish. Sci. Cent. Ref. Doc. 08-02: 35 p.

Wigley, SE. 1996. The Lorenz curve method applied to NEFSC bottom trawl survey data. NMFS NEFSC Ref. Doc. 96-05f. 11 p.