Center for Independent Experts (CIE) Independent Peer Review of the Atlantic Menhaden & Ecological Reference Points Review Workshop

Charleston, South Carolina: November 4-8, 2019

Laurence T. Kell*

*Email: laurie@seaplusplus.co.uk

Contents

Table of Contents

Executive Summary

This Executive Summary provides a concise summary of the findings and recommendations, and discusses whether the science reviewed is the best scientific information available.

The analysts clearly understand the assessment model used, recognise the problems and are addressing these under research recommendations. The meeting benefited from the participation of stakeholders who provided background information and personal experience on the operation of the fisheries and helped with clarifying management objectives.

The current use of single species reference points makes sense as they allow trends to be monitored. However, their robustness needs to be evaluated, and Ecological Reference Points need to be established. Currently, the Multi-Species models are not ready to provide advice, so stock biomass, abundance, exploitation and stock status and reference points from the base run of the BAM should be used for management. Out of the ecosystem models, NWACS-MICE appears to be best able to address management objectives when combined with the single species BAM assessment. For example, the trade-offs between menhaden and striped bass were well explained and presented, e.g., using the NWACS-MICE.

The VADER model may be appropriate for conditioning Operating Models (OMs) in Management Strategy Evaluation (MSE) and for simulation testing BAM and advice based upon it as both BAM and VADER are statistical-catch-at-age models and share similar datasets.

There appear to have been large changes in the life history parameters, and non-stationarity in production processes, i.e. whether the dynamics are driven by the environment or by density dependence. The development of empirical length-based indicators would be a way of monitoring such changes and may help identify plausible model scenarios.

The indices of abundance and catch-at-age appear to have little information in them. Likelihood profiling of key parameters such as virgin biomass or $R0$ by data components for the Base Case could have identified the potential impacts of the different datasets and suggest appropriate sensitivity runs to conduct or uncertainties to try and resolve.

In the single species assessment, uncertainty due to model structure and fixed inputs (e.g., M) is important, due to the lack of information in the abundance indices and potential problems with the catch-at-age data. It would have been valuable to have performed likelihood profiling for the different data components and to have compared estimation with model error. In the former case, this would identify what parameters are influenced by which dataset and help to identify parameter ranges and bounds. In the latter case this could have shown whether uncertainty around point estimates from the Base Case is greater than the uncertainty between sensitivity runs. This is potentially useful as it could have identified whether the data are of sufficient quality, whether alternative hypotheses should be investigated, and the robustness of current assumptions.

For the ERP models, less attention was given to estimation error. Instead the focus was on model error. The comparisons presented in the ERP report generally suggest qualitative agreement across models. However, this alignment is not surprising given the common datasets used to inform the various models and the fact that a main model diagnostic was the comparison of models based on outputs, not their ability to predict observations.

There is also potential for using MSE to evaluate the relative value-of-information versus the value of control, i.e., whether it is better to collect more data or implement more robust management. However, conducting MSE will take several years and require a corresponding commitment of resources, which may result in effort on other tasks being reduced. Therefore, a detailed workplan should be developed where responsibilities and potential benefits are clearly identified.

1 Background

This document contains my independent report of review activities and findings from SEDAR 69 for the 2019 ASMFC Atlantic Menhaden Single-Species Benchmark and Ecological Reference Points Benchmark Peer Reviews, which was held on November 4-8 2019 in Charleston, South Carolina.

Prior to the meeting the ToR for the review panel (RP), the assessment documents, and background material were provided via the cloud and email. The Terms of Reference (ToR) were wide ranging as they covered the single species assessment of menhaden and its role in the ecosystem. Therefore, consideration of both single species and ecosystem-based fishery management was required. Both reports were over 400 pages and 14 presentations were made at the meeting. In addition, I had asked for the assessment inputs and outputs so that it was possible to review the actual assessment itself and to be able to run additional analyses if required.

The expertise of the panel was broad reflecting the ToR and consisted of Dr. Michael Jones (Chair), and Council of Independent Experts (CIE) reviewers Dr. Kenneth Frank, Dr. Laurence Kell, and Dr. Daniel Howell. In addition, although not a CIE reviewer Dr. Sarah Gaichas was a member of the review panel. Dr. Michael Jones is Professor Emeritus at the Quantitative Fisheries Center at Michigan State University. Dr. Kenneth Frank is Research Scientist at Fisheries and Oceans Canada. Dr. Laurence Kell is Visiting Professor in Fisheries Management at Imperial College London. Dr Daniel Howell is Research Professor at IMR, Norway. Dr. Sarah Gaichas is Research Fisheries Biologist at NOAA.

As Chair of the Panel, Dr. Jones facilitated the meeting and made sure that all the ToR were reviewed. He also led the preparation of the Peer Review Panel Summary Report; with Drs. Sarah Gaichas, Daniel Howell, and Kenneth Frank and I serving as independent and impartial reviewers. At the review meeting the members of the ASMFC Atlantic Menhaden Stock Assessment Subcommittee and ASMFC Atlantic Menhaden Technical Committee presented the Stock Assessment and Ecological Reference Point Stock Assessment Reports. The meeting also benefited from the participation of stake- holders who provided background information and personal experience on the operation of the fisheries and helped with clarifying management objectives. The structure of the meeting was fairly informal with discussion during each presentation, which allowed a lot of interaction that I was able to actively participate in with the other members of the RP.

Following the meeting, the CIE reviewers each completed independent peer review reports in accordance with the requirements specified in the Statement of Work and ToR (Appendix A). In adherence with the required formatting and content guidelines; reviewers were not required to reach a consensus. As a CIE Reviewer, I submitted my Individual Peer Review Report and contributed to the Peer Review Panel Summary Report. During the meeting there was a general consensus among the RP regarding most of the main discussion points and findings which are outlined in the Panel Report.

2 Reviewer Role

The ToR covered both single species and ecosystem benchmarks. My experience covers data poor and rich stock assessments, model validation, and Ecosystem Based Fisheries Management (EBFM). I also have an interest in the development of robust management advice given uncertainty and the development of precautionary and target reference points using Management Strategy Evaluation (MSE). This is of particular importance since a range of models with different assumptions and data requirements were used and there are multiple management objectives with potential trade-offs between them, e.g., related to use of resource by the industry, the supply of ecosystem services, and sustainable use.

After the Review Workshop, I assisted in finalising the Workshop Report and prepared my independent report. As stated above, the stock assessment inputs and outputs were provided by the assessment team (AT) on request and this helped me in my review, especially as it allowed me to conduct sensitivity analyses and to evaluate uncertainty. I felt I was able to make a contribution by helping to propose a variety of diagnostics and additional analyses, and how MSE could be conducted in the future. I worked with the panel to finalise the review of the ToR and helped to draft the report. In this report, I do not simply repeat the conclusions set forward in the Summary Report, with which I agree, rather it focuses on areas where I could make the most significant contribution.

3 Terms of Reference

Under this section I summarise the work presented at the review with respect to the ToR, identifying strengths and weaknesses. I expand on the areas where I feel my experience is of most relevance. This is primarily related to single species assessment and management. There are links between the single species assessment and the ecological reference points (e.g., model validation, development of robust management advice and Management Strategy Evaluation), which I also cover in the Ecological Reference Point section and then expand on in the Recommendation and Conclusions section.

3.1 Terms of Reference for the Atlantic Menhaden Single-Species Peer Review

ToR 1: Evaluate the thoroughness of data collection and the presentation and treatment of fishery-dependent and fishery-independent data in the assessment

The main datasets used in the assessment were time series of indices of abundance, catch-at-age and growth and fecundity. As there are no fishery independent surveys targeting menhaden, a number of surveys are combined into regional adult abundance indices and a young of the year (YOY) index. Three adult (i.e., Age-1+) indices were developed from eight fishery independent survey data sets: Northern (NAD; age-2+); Mid (MAD; age-1+); and Southern-Atlantic (SAD; age-1), while the YOY index was developed from 16 fishery-independent surveys.

Data on growth and fecundity are also collected and tagging data have been used to estimate natural mortality (M). These life history parameters are required as inputs into the age-based assessments, for the estimation of priors, and for reviewing changes in stock productivity.

The 16 indices used to develop the YOY index were reduced from an original total of 49 using a set of criteria; namely i) absence of hyperstability or gear saturation; ii) length of time series (minimum 10 years); iii) spatial extent; iv) inclusion of trips with zero catches; v) consistent data collection over time;
vi) identification of catches to species level; v) convergence of the model used for standardization; vi) whether information on gear selectivity is available to determine if the index is for YOY or adult fish; and vii) availability of length data.

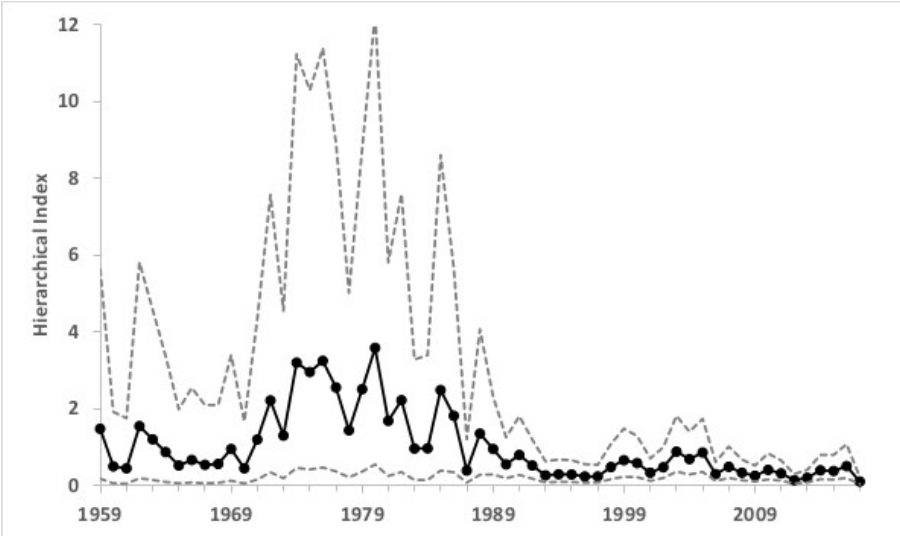

Figure 108. Time series of the young-of-year (YOY) Atlantic menhaden relative abundance index as estimated from hierarchical analysis (Conn 2009). The black line gives the posterior mean and the grey, dashed lines represents a 95% credible interval about the time series.

Figure 1: YOY index as presented in stock assessment report.

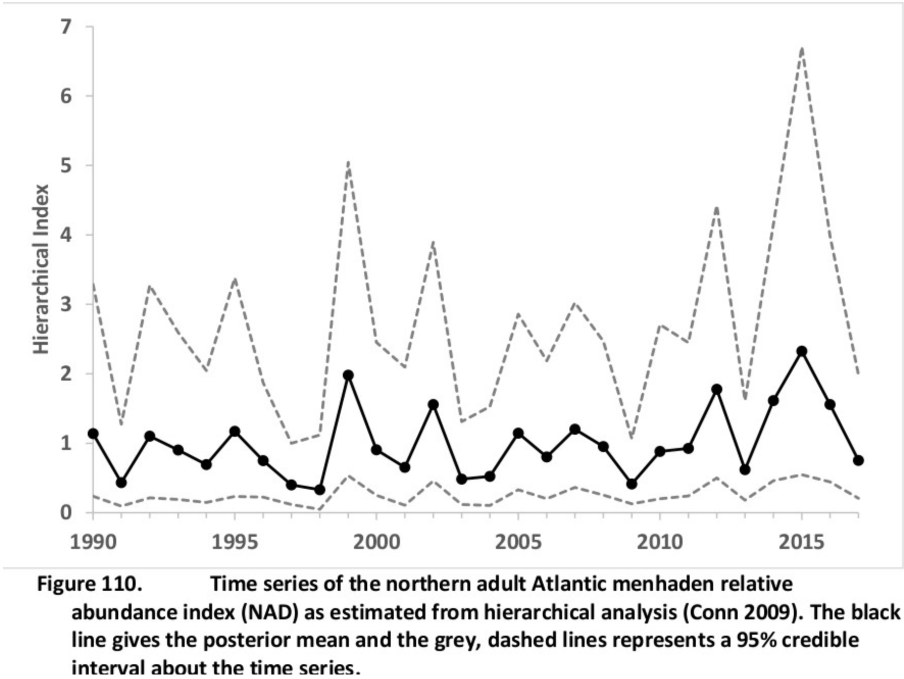

Figure 2: NAD index as presented in stock assessment report.

The YOY index was then combined coastwide using hierarchical modeling (Conn (2009)). This assumes each index samples relative abundance, subject to sampling and process errors. The former is the error attributed to within-survey variance or variation in methodology (field conditions, observer experience, human error); while the latter is the error attributed to variation in catchability, spatial distribution, variation due to biotic or abiotic processes due to temporal variation in index-specific catchability (and possibly to differences in selectivity between gear types). It is also assumed that the indices are measuring the same relative stock abundance and that the surveys have similar selectivities. Choice of regional structure was based on length frequency analyses, which indicated differences in gear selectivity and age-specific seasonal migrations of Atlantic menhaden.

The calculation and standardization of the main abundance indices was thorough and justification for inclusion or elimination of the available data sources was made considering the data strengths and weaknesses. Standardised indices were presented with confidence intervals, e.g., Figures 1 and 2 for YOY and NAD. This allowed factors related to the number of surveys and differences in sampling programs to be discussed. The approach took into account uncertainty in both temporal and spatial scales, gear selectivities, aging accuracy, and sample size. The indices, however, appear to have little contrast; given the absence of surveys targeted at menhaden, this is an appropriate approach.

The indices of abundance are used for calibration of stock assessment methods. If indices are in conflict or uncorrelated then model estimates may be uncertain or biased. Therefore, the correlations between indices may suggest alternative data weightings and/or scenarios to run. Therefore, a first step, before fitting stock assessment models, is to identify whether indices are in conflict. Therefore, the indices are plotted against each other and the correlations calculated in Figure 3.

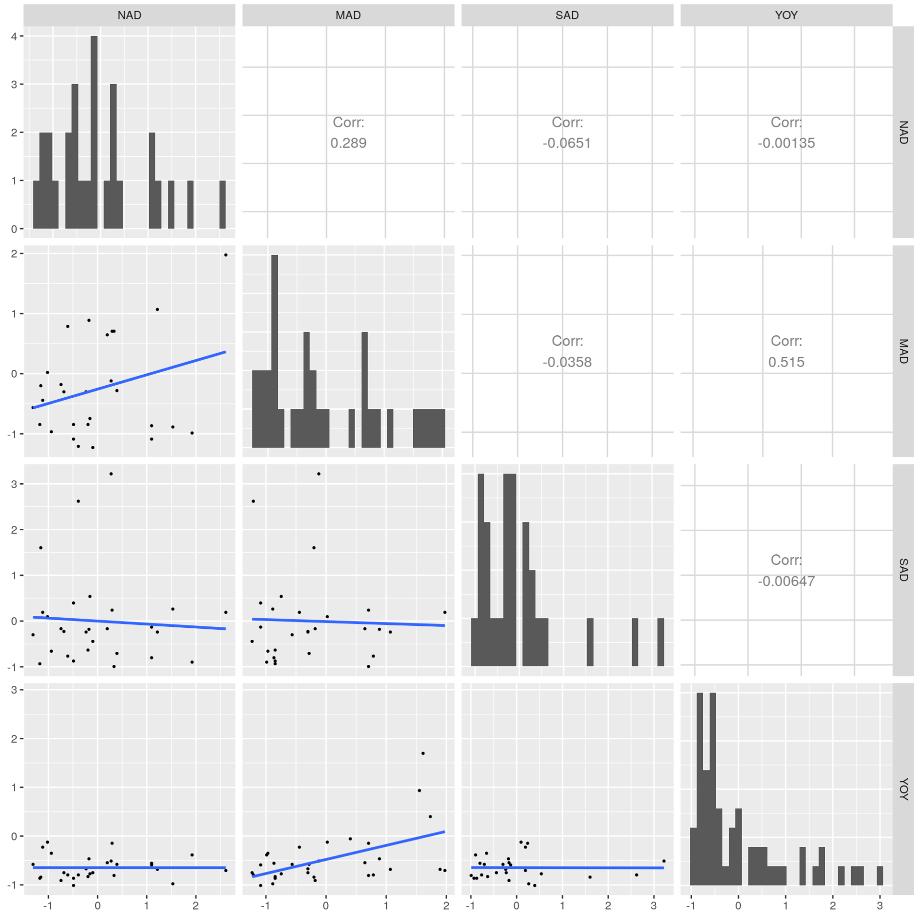

Figure 3: Correlations between the indices used in the BAM assessment.

There appears to be no or only weak correlations between the indices, i.e., NAD v MAD and MAD v YOY. This lack of correlation is likely to result in large uncertainty in parameter estimates and differences between scenarios using alternative datasets, model structures or ways of estimating parameter uncertainty.

A reason for poor correlation may be because indices represent different age classes or due to spatial heterogeneity. In the former case, there may be lags between indices, which can be evaluated by plotting cross-correlations, e.g., if there are cohort effects on abundance the YOY may peak prior to that of an adult index. The cross correlations between the indices are therefore plotted in Figure 4, in which there appears to be a lag between YOY and MAD which indicates that the two indices may be tracking the same cohorts.

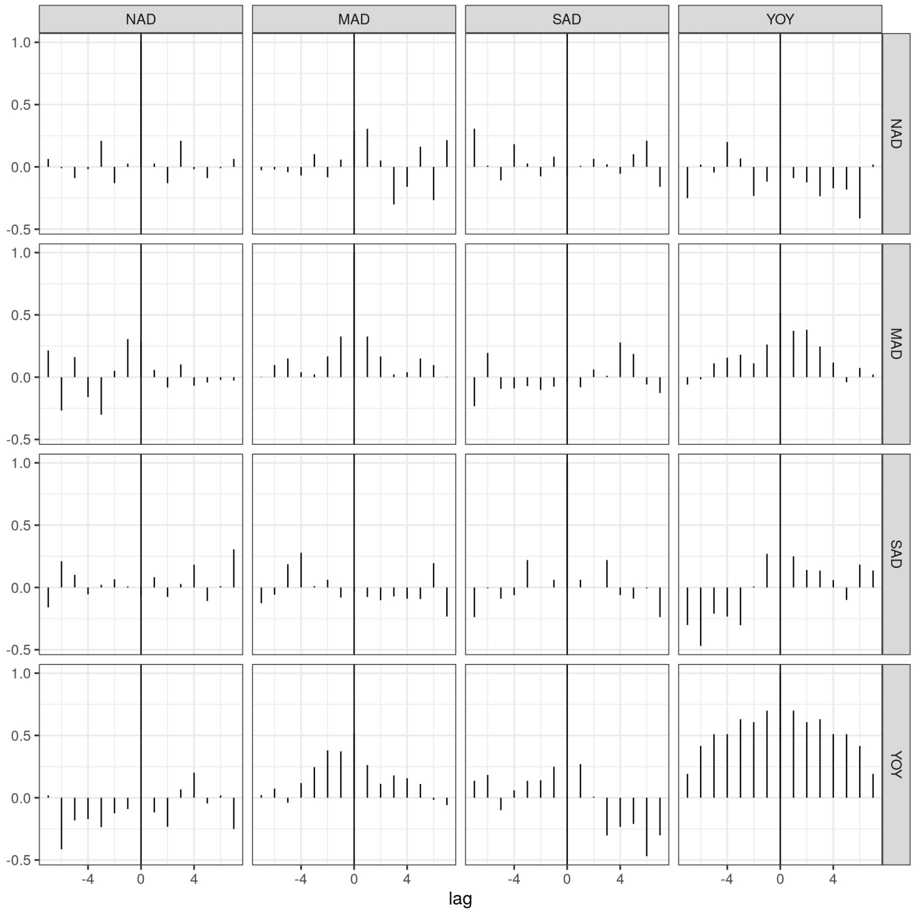

Figure 4: Cross correlations between indices.

While indices of abundance are important for estimating stock productivity, catch-at-age data are important for tracking cohorts in the population. Therefore in Figures 5 and 6 the catch-at-age for the reduction and bait boat fishery respectively are plotted; these are scaled by the mean across ages to allow year-class effects to be examined. There does not seem to be much information in these data on cohort strength. Also, high abundances appear to occur in adjacent years which could indicate aging problems.

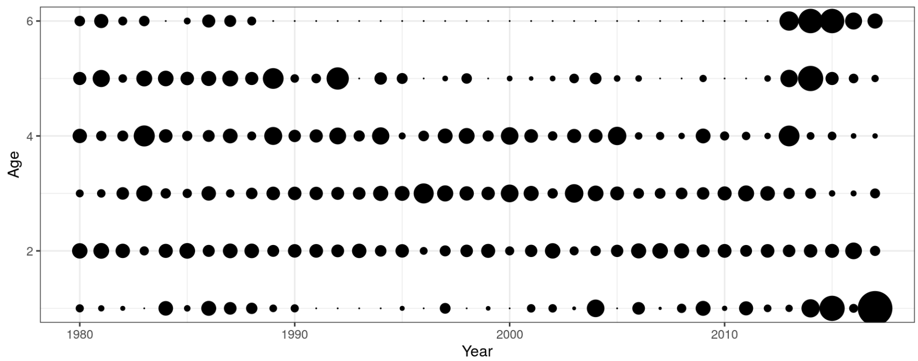

Figure 5: Catch-at-age of reduction fishery, ages scaled by mean.

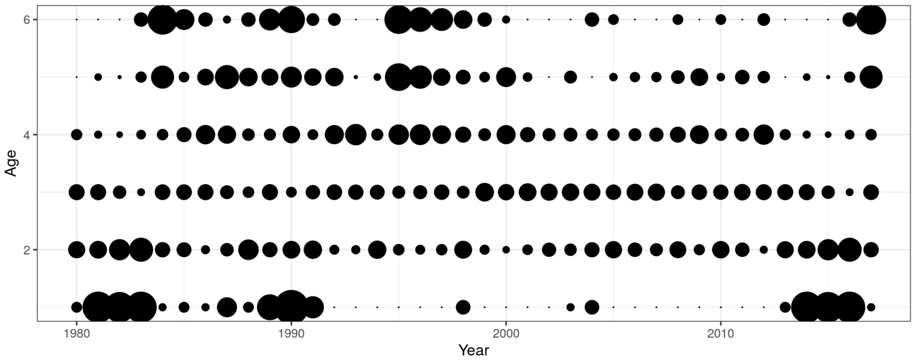

Figure 6: Catch-at-age of bait boat fishery, ages scaled by mean.

Figure 7 shows the length composition of the NAD index plotted with the length composition of fishery samples (pooled over all regions). This shows a strong residual pattern and suggests that a dome- shaped selectivity may have performed better for that survey. However, this would leave the model with no logistic selectivity tuning datasets and hence no direct data constraint on the modelled number of larger fish. The panel concluded that the logistic selection for the NAD was the appropriate choice at present, but recommended a re-evaluation of the available data aimed at identifying a series that provided better coverage of the larger fish, and in the longer term gave a recommendation to survey these larger fish directly.

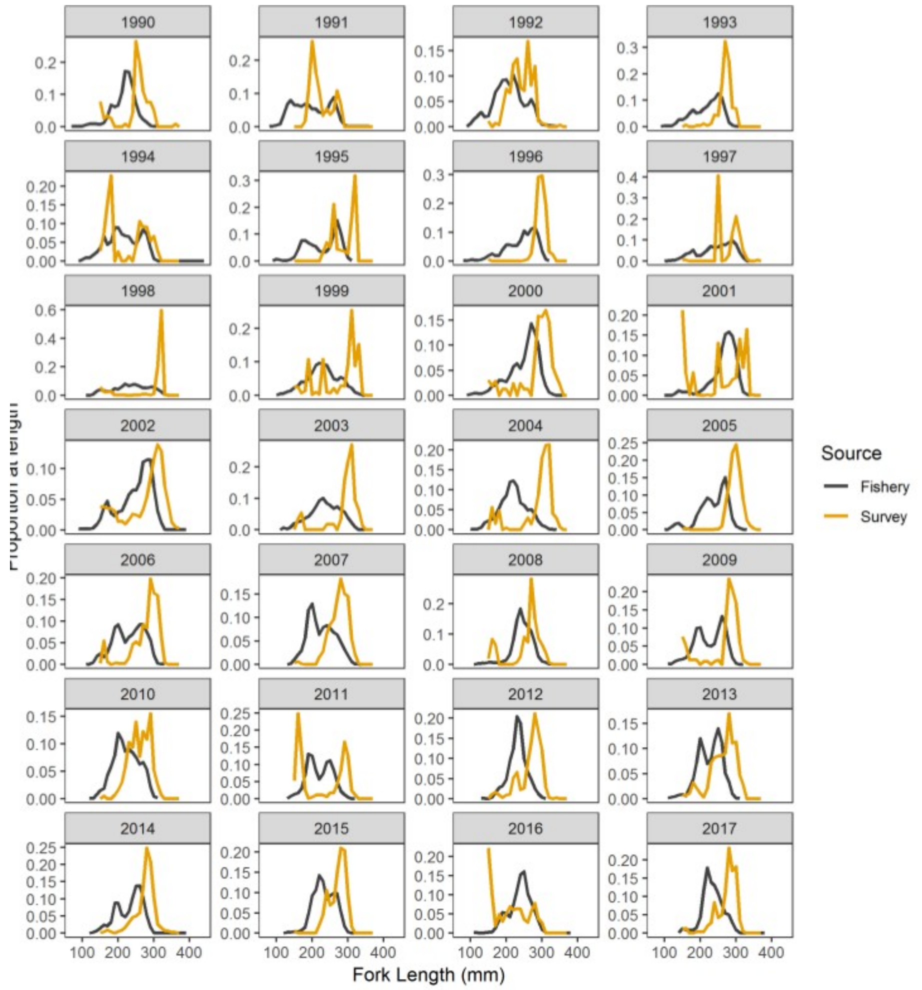

Figure 113. Length composition of the Northern Adult (NAD) index plotted with the length composition of fishery samples (pooled over all regions).

Figure 7: length composition of the NAD.

There have been large changes in biological parameters, e.g., fecundity (Figure 8). Life history parameters are summarised in Figure 9 and the correlations between them in Figure 10.

Major changes have been seen in the life history parameters. For example, there has been an increase in k, a decrease in L_∞, also the proportion of age 2 that were mature was low in the 1980s when k was at a low. There appears to have been large changes in the life history parameters, while the indices of abundance and catch-at-age appear to have little information in them.

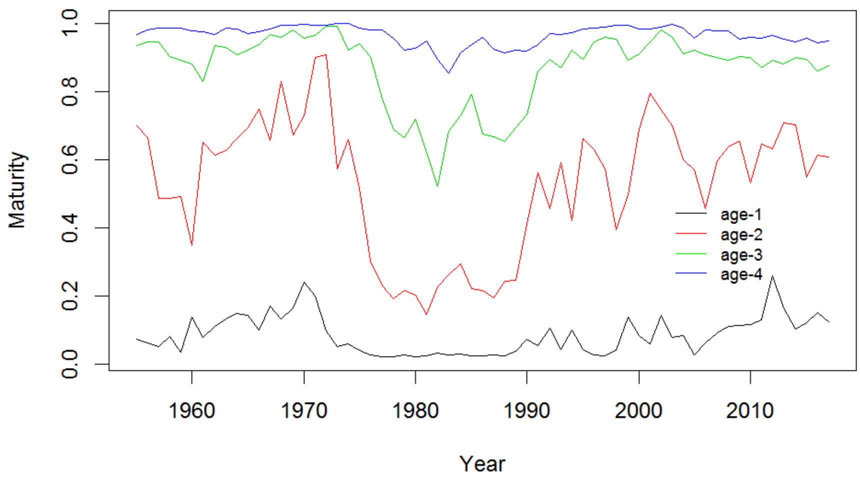

Figure 8: Atlantic menhaden maturity by year and age. Age-0 menhaden are immature, while fish of age-5 and older are 100% mature.

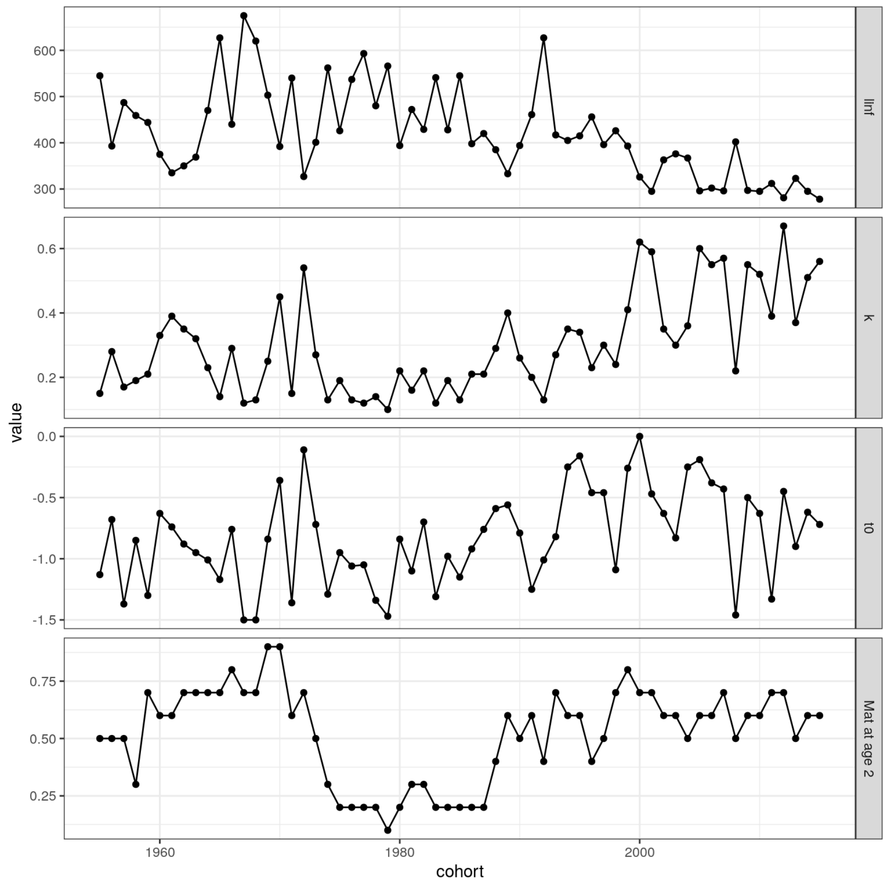

Figure 9: Time series of von Bertanffy growth parameters and proportion mature at age 2.

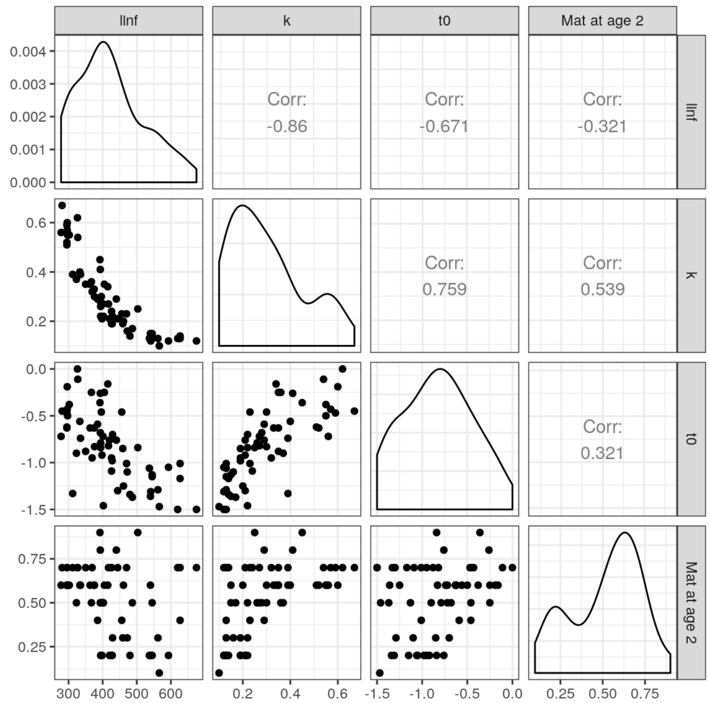

Figure 10: Pairwise scatter plot of Bertalanffy growth parameters and proportion mature at age 2.

ToR 2: Evaluate the methods and models used to estimate population parameters (e.g., F, biomass, abundance) and biological reference points

The Atlantic menhaden assessment was conducted using the Beaufort Assessment Model (BAM), a Statistical Catch-at-Age Assessment (SCAA). Although BAM was the main tool used to provide single species assessment advice, a biomass-based assessment model (SPMTv) which estimated changes in population growth rate (*r*) was applied in the ERP report. BAM as a statistical catch-at-age model estimates population numbers and fishing mortality-at-age and then projects the population forward in time. The main datasets used are indices of abundance for calibration, catch-at-age to estimate cohort dynamics and scaling and life history parameters to estimate productivity and fecundity. BAM has been used in both the previous 2015 benchmark stock assessment and the 2017 stock assessment update. BAM was configured as a fleets-as-areas model where each of the fleets was broken into areas to reflect differences in size and age structure along the coast. The model was fitted to catch-at-age and indices of abundance, both fisheries-dependent and independent. Fishery-dependent data came from the commercial reduction and bait fisheries.

BAM as a single-species Statistical Catch At Age (SCAA) model is an appropriate tool for assessing the Atlantic menhaden stock. In general, model results were well presented, with fits to the base case presented as estimated values with error bars. Residual plots were included where appropriate for indices, age and length compositions and selectivities. These plots provided a basis for evaluating the base case model fit.

As a statistical catch-at-age model BAM, estimates selectivity and time varying selectivity blocks were chosen and justified. Although the choice of blocks could have been simplified, the current scheme represented a viable basis for advice and provided a Base Case against which alternative model scenarios could be compared. The choice of logistic rather than dome shaped selectivity for the NAD index is a potential problem.

Large changes in growth and fecundity have been seen (see above). For example, a comparison of the estimated annual cohort-based von Bertalanffy growth coefficients showed that the average value of k was about 0.2 in the 1950s and is now around 0.5. This is important as Gislason et al. (2008) showed a significant relationship between M and k. It is likely therefore that M is to have varied between years; changes in k are likely to have a large effect on productivity and hence reference points Jensen (1997).

The rationale used for selecting the Base Case was justified and the analysts clearly understand the assessment model used, recognise the current problems, and are addressing them under research recommendations.

ToR 3: Evaluate the diagnostic analyses performed

The main diagnostics presented were residual fits for the base case, sensitivity runs, and retrospective analyses where model outputs were compared.

Goodness of fits diagnostics based on residuals were only presented for the Base Case, while comparison between the Base Case and the sensitivity runs was based on model outputs, e.g., biomass and fishing mortality. This makes it difficult to objectively choose between or to weight scenarios. A reason for using model outputs is because when alternative assessment model structures are developed using different datasets, conventional model selection criteria such as AIC cannot be applied, as in this case where datasets are excluded or BAM is compared to a surplus production model which does not use catch-at-age data.

Is important to try and identify potential model misspecification, i.e., to use diagnostics that can identify conflict between fits to alternative model structures and datasets, and to allow scenario hypotheses to be rejected or weighted. As well as residual plots, there are a variety of diagnostics, including runs tests, likelihood component profiling, comparison between production and age-based models, and hind-casting. Carvalho et al. (2017) found that no single diagnostic worked well, however, and recommended the use of a carefully selected range of diagnostics.

Sensitivity runs for BAM included hypotheses related to input data, changes to the model configuration, and values of fixed parameters. These sensitivity runs were then used in the retrospective analyses.

Figures 11, 12, 13, 14 show a range of sensitivity runs. From Figure 11 it appears that the NAD index is having the biggest effect on fecundity estimates, since when it is removed the level changes, however, trends and peaks remain similar. Bigger differences between runs are seen for the ERP scenarios, where the fishery dependent indices are used.

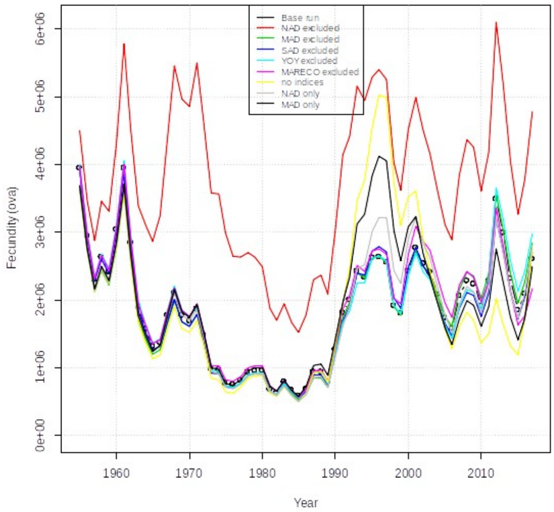

Figure 11: Fecundity in billions of ova for 1955-2017 for a suite of sensitivity runs that explored inclusion and exclusion of indices (Single species report Figure 160).

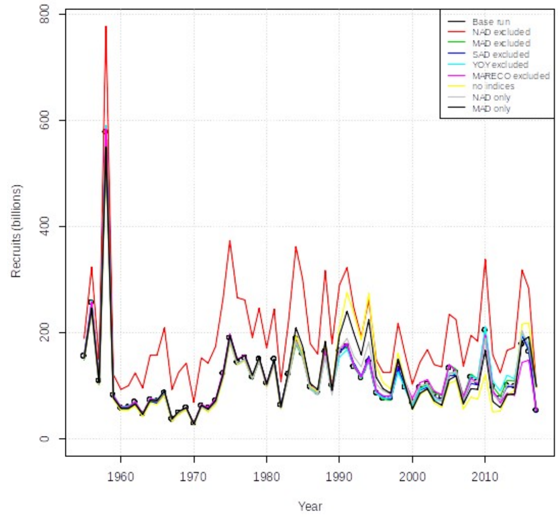

Figure 12: Recruitment in billions of fish for 1955-2017 for a suite of sensitivity runs that explored inclusion and exclusion of indices (Single species report Figure 161).

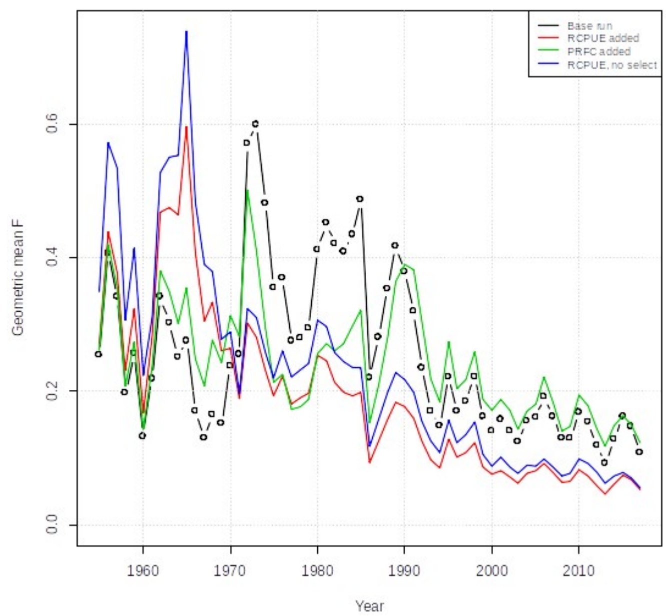

Figure 13: Geometric mean fishing mortality rate for ages-2 to -4 for 1955-2017 for a suite of sensitivity runs that explored inclusion and exclusion of indices from the work being completed by the ERP group (ERP Report, SEDAR 2019) (Single species report Figure 163).

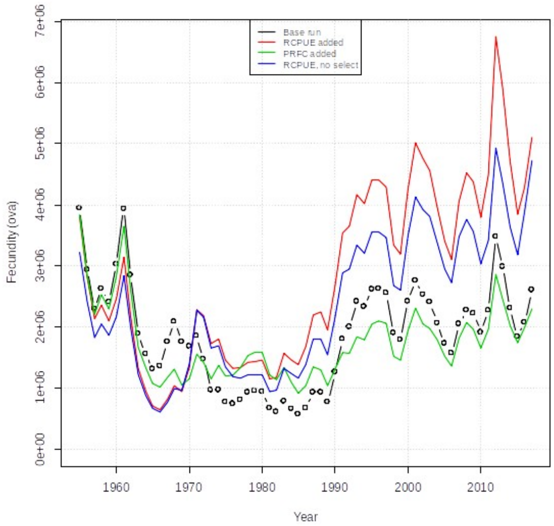

Figure 14: Fecundity in billions of ova for 1955-2017 for a suite of sensitivity runs that explored inclusion and exclusion of indices from the work being completed by the ERP group (ERP Report, SEDAR 2019) (Single species report Figure 164).

An example of the use of runs test to evaluate whether the data are randomly distributed around a central tendency without any systematic patterns, is shown in Figure 15.

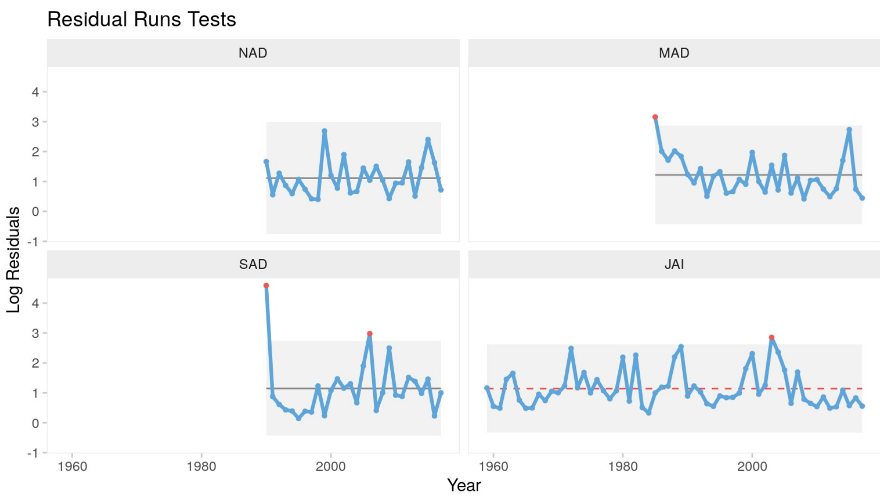

Figure 15: Runs chart for base scenario showing the residuals by year; red points indicate points that violate the 3-sigma rule, and the red dashed line indicates unusually long runs or unusually few crossings.

Likelihood profiling of key parameters such virgin biomass (K) or recruitment ($R0$) by data components, for the Base Case could have identified the potential impacts of the different datasets and suggested appropriate sensitivity runs to conduct or uncertainties to try and resolve.

It is important to be able to compare model fits across sensitivity runs in an objective way. There are several approaches that have potential, e.g., residual runs tests, likelihood profiling of data components (Carvalho et al., 2017) and hindcasting (Kell et al., 2016) that could be considered in the future.

ToR 4: Uncertainty in Estimated Parameters

Uncertainty in stock status was estimated using two approaches, the bootstrap (MCB) to estimate estimation error and a MCMC analysis to evaluate the impact of uncertainty in inputs.

There was little contrast in the abundance indices, e.g., Figure 2 which shows the northern adult Atlantic menhaden relative abundance index (NAD) and Figure 1 for the young-of-year index (YOY). The black line gives the posterior mean and the grey, dashed lines represent a 95% credible interval about the time series. A straight line could be drawn through these indices. In the BAM, the input M is larger than the estimates of F and there is little cohort signal in the catch-at-age, therefore age based dynamics will be difficult to estimate with any precision.

Uncertainty due to model structure and fixed inputs (e.g., M) is important, since the lack of information in the abundance indices and potential problems with the catch-at-age data. It would have been valuable to have performed likelihood profiling for the different data components and to have compared estimation with model error. In the former case, this would identify what parameters are influenced by which dataset and help to identify parameter ranges and bounds. In the latter case this could have shown whether uncertainty around point estimates from the Base Case is greater than the uncertainty between sensitivity runs. This is potentially useful as it could have identified whether the data are of sufficient quality, whether alternative hypotheses should be investigated, and the robustness of current assumptions.

The tuna Regional Management Fishery Organisations (RFMOs) commonly base advice on an uncertainty grid where estimates from multiple scenarios are compared rather than estimation error from a Base Case (e.g., Kell et al., 2015). It will be time consuming to estimate uncertainty using MCMC and MCB, However, uncertainty could be derived from the covariance matrix. While in previous assessments, it was thought that these were an underestimate of uncertainty, they could provide a useful comparison with other methods and are quicker to run than MCB and MCMC, and so could be done for all scenarios.

Sensitivity runs and retrospectives were conducted, and the main diagnostic was to compare model outcomes, i.e., estimates of SSB and F. A fuller exploration of diagnostics, e.g., runs tests and hindcasting, could have identified potential model misspecification.

The treatment of uncertainty and how it propagates through into advice is important. However, forecasts were not included in the ToRs. However, this is dealt with in ToR 7 below.

ToR 5: If a minority report has been filed, review minority opinion and any associated analyses

The panel was in agreement about the quality of the work presented at the workshop, and so no minority report was filed.

ToR 6: Recommend best estimates of stock biomass, abundance, and exploitation from the assessment for use in management, if possible, or specify alternative estimation methods

A concern with the BAM was whether solutions are from a global optima, as during the meeting it was found that the optimizer sometimes failed to converge. It is important therefore to conduct jitter analyses for all runs, including sensitivities and retrospectives.

The problem of uncertainty can be particularly acute when more than one data source is available, and these are potentially in conflict as different likelihood weightings for the different data components may provide contradictory parameter estimates. The proposed likelihood weighting scheme was considered appropriate, although time constraints prevented a more in-depth analysis of the weighting scheme. Likelihood profiling of key parameters by data component for the Base Case would be a useful exercise in future.

Alternatives to BAM are Stock Synthesis 3 (Methot, 2005, SS3) and surplus production models. SS3 would have allowed length compositions for each index to be included in the assessment (e.g., to allow ageing error to be evaluated) and the impact of alternative weightings scenarios for each index and data types (i.e., indices, abundance and length compositions) to be evaluated. Although surplus production models were presented in the context of the ERP work, these were not considered to represent a viable alternative to the BAM model. A reason for this was the lack of contrast in the indices of abundance which would make it hard to estimate the production function. There is, however, an implicit production function in BAM, which due to the lack of contrast in the CPUE data is difficult to estimate, and so may be mainly influenced by the assumed life history parameters. A potential approach to evaluate variation in life history parameters and selectivity is to use JABBA-Select Winker et al. (submitted) a surplus production model that incorporates life history parameters and fisheries selectivity and distinguishes between exploitable biomass and spawning biomass, enabling more direct comparison with age-structured model results.

A main problem is lack of contrast and information in the data, and as the analysts are experienced with the BAM, then the use of a single model is not considered a problem.

ToR 7: Evaluate the choice of reference points and the methods used to estimate them. Recommend stock status determination from the assessment, or, if appropriate, specify alternative methods/measures.

Current fishing mortality reference points are based on the fishing mortality rates estimated by BAM for ages 2 to 4 during the period 1960-2012, while the target is estimated as the median and the threshold by the maximum geometric mean respectively. The reference points for biomass are based on reproductive output and are estimates by population fecundity (FEC, number of maturing or ripe eggs) a measure of reproductive capacity. The biomass reference points are the FEC values associated with the fishing mortality target and threshold predicated on calculations for spawning potential ratio (SPR). All benchmark calculations were based upon landing selectivities weighted across all fleets and areas, M-at-age (which was constant), mean maturity-at-age, a 1:1 sex ratio, and mean fecundity-at-age used as inputs to BAM, estimated across the entire time series from 1955 to 2017. It is reasonable not to use $FMSY$ reference points for a forage species that provides ecosystem services. It was also agreed that the use of a reference period when the fishery was open access without management constraints was appropriate.

Inspection of the life history parameters (i.e., Figures 9 and 10) show that these have not been constant over time. For example, Figure 8 shows maturity by age, which in the mid- to late 1970s showed a sharp decline (i.e., from 80% to 20%), recovering to vary around a level of about 60%. Given the large changes seen in life history parameters, the robustness of the time periods chosen should be evaluated, particularly compared to future conditions.

Although M is considered to be time invariant, it is likely that it will have varied over the time series, and as BAM estimates F to be less than M, this will have an important impact on historical estimates, the perception of the impact of fishing, and reference points. Basing reference points on past stock and fishery performance is reasonable, however, since the level of fishing observed during the reference period provides comfort that stock will not collapse if fishing is maintained at same level. The year ranges used in the estimation of reference points should be evaluated to ensure that they show robustness to the different sources of uncertainty.

How uncertainty is propagated into probabilities is important as the reference points are based on medians and maximum values. Uncertainty in the benchmark was estimated in two ways, by the bootstrap (MCB) to estimate estimation error and an MCMC analysis where the biological parameters were resampled. For each MCB run and MCMC analysis, reference points were re-calculated, allowing a distribution of the benchmarks to be derived, see Figure 16 and 17.

Although the expected values are similar, the distributions are different. For example, fishing mortality distributions are bimodal for the BAM MCB runs. This may be due to non-convergence, hitting bounds, or other problems with the assessment. The strong correlations in the MCB runs suggest that there is little information in the data to estimate population parameters, i.e. whether the stock is large or highly productive. In contrast, the correlation and spread for the MCMC runs is less. The consequences for advice need to be investigated further.

Single species reference points make sense as they allow trends to be monitored. However, their robustness needs to be evaluated and Ecological Reference Points also need to be established.

Forecasts

Evaluation of stock forecasts were not included explicitly in the ToR, however, a main reason for conducting stock assessment and estimating reference points is to set management measures such as total allowable catches. Due to the assumed high level of natural morality, the stock biomass is likely to driven by incoming recruitment. There does not appear to be a stock recruitment relationship. However, recruitment has been relatively stable with no sign of recruitment failure and variability in recruitment is not exceptionally high (CV of 30%). In the projections, a non-linear time series (NLTS) approach was used to model incoming year-classes, which, given the lack of a stock recruitment relationship and relatively stable recruitment, is appropriate.

The ability of the stock assessment model to forecast the future state of the resource is important if management advice is to be robust. Currently, there is a lag between the last year in the assessment and the year for which the TAC is set, e.g., the last year in the current assessment is 2017 which will be projected for reported landings in 2018 and preliminary estimates of 2019 catches, and the TAC will then be set for 2020 to 2022. Therefore, an evaluation of prediction skill is important. This could be performed by conducting a hindcast. This is similar to a retrospective analysis where the most recent years in the assessment are removed and the stock projected for the reported catches. The performance of the forecast can then be compared with the historical outcomes (Kell et al., 2016).

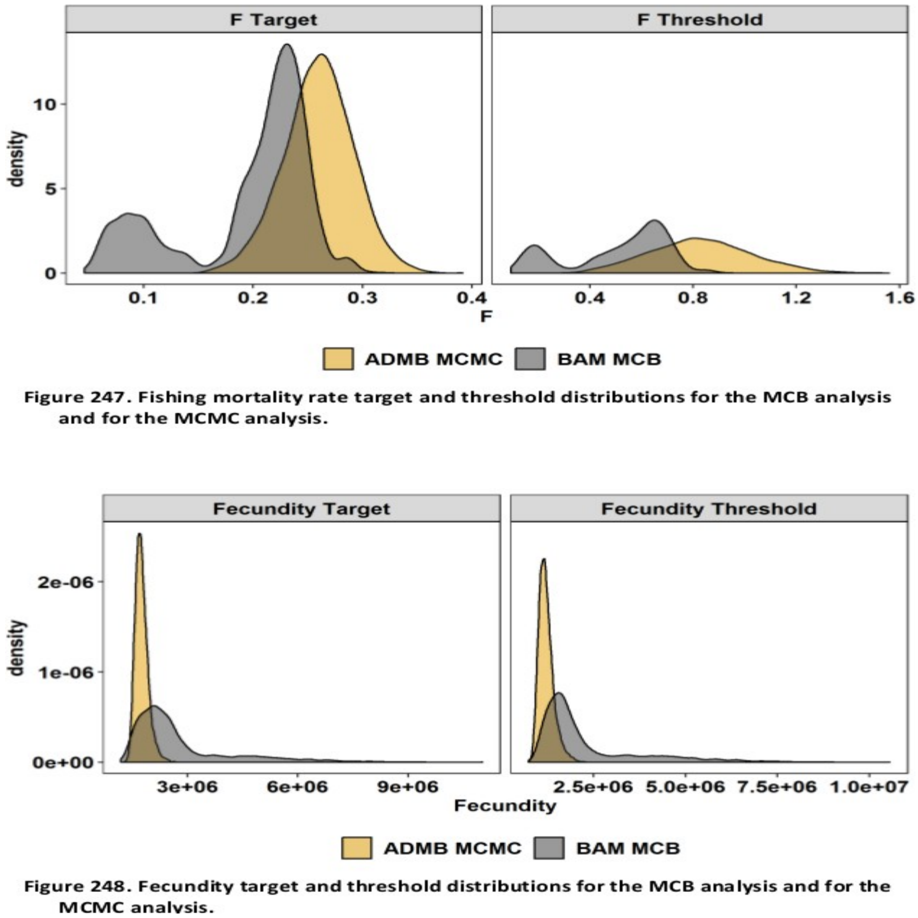

Figure 16: Probability distributions of Atlantic menhaden reference points.

Figure 18 plots SSB against fishing mortality for the BAM base case estimates. An increase in F would be expected to result in a decrease in SSB, and vice versa if the dynamics are driven by a production function. However, this does not seem to be the case, as taking the first point in 1955 the path moves clockwise. This implies that dynamics are driven by process error, i.e., changes in growth or recruitment and not fishing.

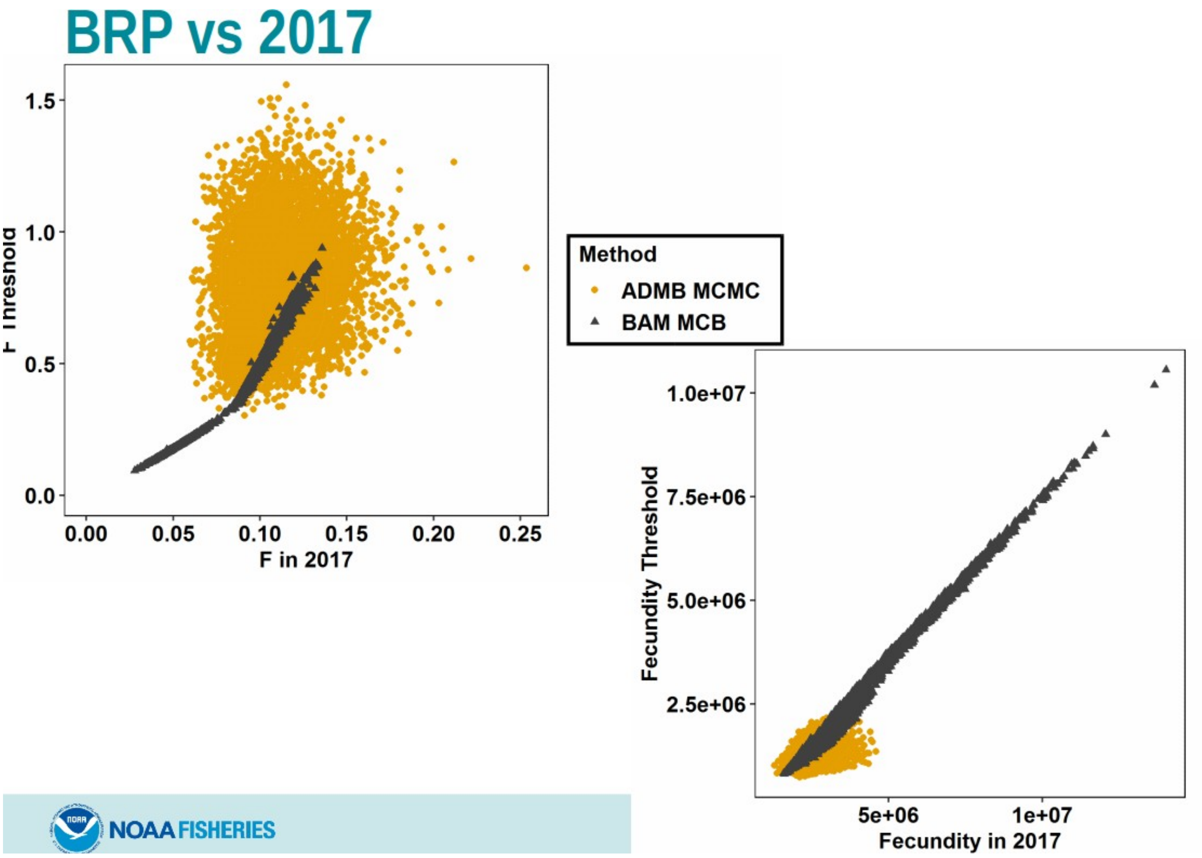

Figure 17: Bi-plots of Atlantic menhaden reference points.

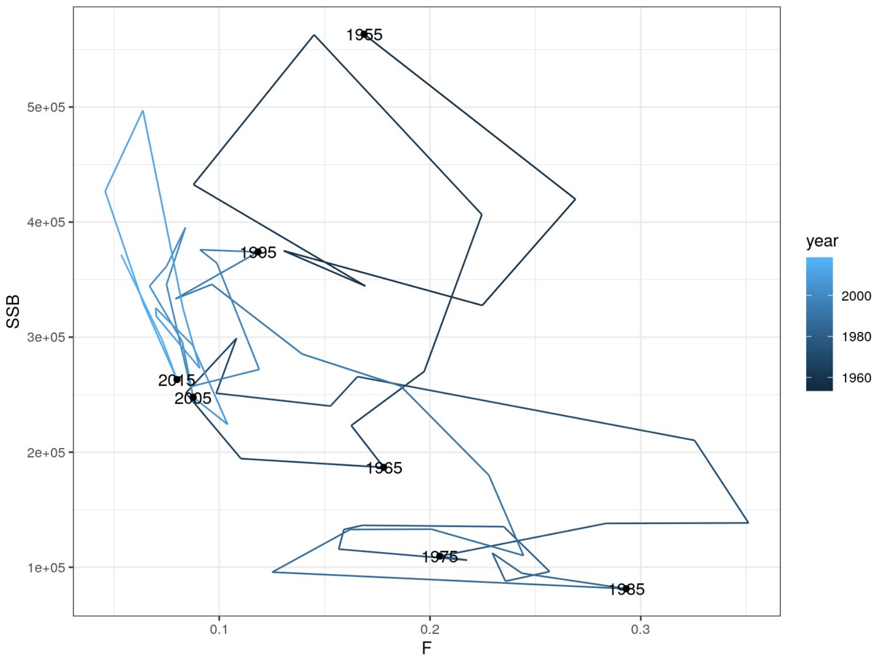

Figure 18: Phase plot of SSB against F for the BAM Base Case.

This has various implications. Assuming the model is not misspecified, is it possible to make forecasts? A way of estimating the ability of a stock assessment to provide forecasts is to use estimates of surplus production from an assessment model (Walters et al., 2008). This can provide a check on whether predictions of changes in biomass ($B_{t+1}B_t$) can be made reliably based on catch and current biomass, in particular whether similar B levels have exhibited similar SP at different historical times, i.e., whether there has been non-stationarity in production processes, whether the dynamics are driven by the environment or by density dependence.

ToR 8: Review the research, data collection, and assessment methodology recommendations provided by the TC and make any additional recommendations warranted. Clearly prioritize the activities needed to inform and maintain the current assessment, and provide recommendations to improve the reliability of future assessments.

There are concerns about the collection, assembly and treatment of the fishery-dependent data from the commercial reduction and bait fisheries. Between 1955-1984 bait landings reporting is possibly incomplete, although from 1985 to present the data are more reliable due to improvements made to the harvester and dealer reporting program. Sampling of bait fishery for length and age has improved since 1988. Expanded sampling was recommended in the bait fishery given the deficiency of age-5+ fish and the top of hold sampling in the reduction fishery should be examined further to ensure accurate characterization of the total trip catch, not just from the last tow which is current procedure. A major concern is the lack of data on the large fish. Therefore collecting better data on the larger fish is identified as the most pressing data collection recommendation for this stock.

There is also potential for using MSE to evaluate the relative value-of-information versus the value of control, i.e., is it better to collect more data or implement more robust management.

ToR 9: Recommend timing of the next benchmark assessment and updates, if necessary, relative to the life history and current management of the species

I agree with recommendations in the Panel Report.

3.2 Terms of Reference for Atlantic Menhaden Ecological Reference Points Peer Review

ToR 1. Evaluate the justification for the inclusion, elimination, or modification of data from the Atlantic menhaden single-species benchmark assessment.

The fishery independent indices of abundance used in BAM lack contrast and so provide little information for estimation of production functions. Therefore, two additional long-term fishery-dependent indices of abundance for Atlantic menhaden were considered for use in the ERP assessments. These were the commercial reduction fishery CPUE index (RCPUE index) and the Potomac River Fisheries Commission index (PRFC) derived from the commercial bait fishery.

Figure 19 shows the lack of signal in the fishery independent indices and conflicts in the fishery dependent indices.

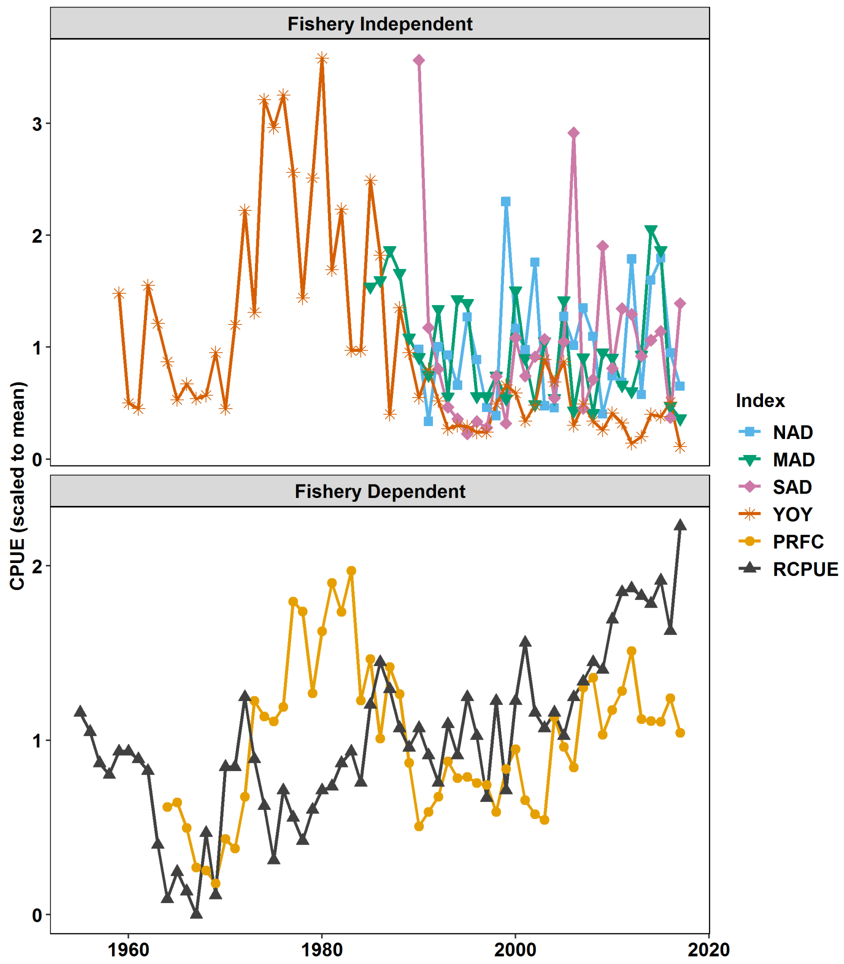

Figure 19: Fishery independent (top) and fishery dependent (bottom) indices of abundance for Atlantic menhaden (Figure 13 from ERP report).

Since the two commercial indices showed different trends, the ERP WG elected to use the RCPUE index for the base case runs because of its larger spatial coverage, consistently recorded unit of effort, known variance structure, support from supplemental analyses that showed relatively strong correlations with other sources of data, and the ability to standardize the data through explanatory covariates (e.g., week, factory, vessel size). While this is reasonable, the cause of the difference between the two fisheries could be important for management for example if there has been a change in the operation of the fleets or distribution of the stock.

The single species assessment input datasets were used for the multi-species SCAA model, and BAM outputs were used for tuning the EwE models. While this ensured consistency, it also means that the true uncertainty may be underestimated, given the potential model misspecification in the single species assessment and the conflicts in the indices. An alternative would be to allow the ecosystem models to estimate biomass, F, and other quantities of interest and comparing these to the single species model. The problem remains of how to objectively validate the models.

ToR 2: Evaluate the thoroughness of data collection and the presentation and treatment of additional fishery-dependent and fishery-independent data sets in the assessment

Although I am not an expert on data collection for ecosystem models, I believe the data collection, and treatment of fishery-dependent and fishery-independent data sets required for the ERP models were both thorough and appropriate.

The strategy adopted of using datasets, i.e., fishery independent indices, total catch, and fishery dependent and independent age and length data, directly from previously reviewed assessments for use in the ecosystem models streamlines the process by relying on existing review processes.

ToR 3. Evaluate the methods and models used to estimate Atlantic menhaden population parameters (e.g., F, biomass, abundance) that take into account Atlantic menhadens role as a forage fish

This ToR was interpreted as meaning to evaluate the ERP models used to estimate reference points, not the predator stock assessments that provided inputs to the ERP models.

Five models were used, ranging in structural complexity and data requirements from a surplus production model with time-varying population growth rate (SPM-TVr) to a full food web model for the Northwest Atlantic continental shelf (NWACS-FULL). Intermediate models included a surplus production model that accounted for menhaden predation (Steele-Henderson), a multispecies statistical catch at age model with menhaden and five species (VADER), and a scaled down food web model (based on NWACS-FULL) that only included menhaden and a subset of key predator and prey species (NWACS-MICE).

The ERP report provided a clear summary of the performance and requirements of each model, and models outputs were provided in comparable forms (e.g., age 1+ biomass and exploitation rate).

The model selected to provide advice was NWACS-MICE, which uses the equilibrium Ecopath model to initialise the dynamic Ecosim model, which is then calibrated using sum of squares fits to time series of biomass and catch for multiple species. There is therefore a degree of circularity; objective ways to compare and validate such models is a worthy goal.

This is a cut down version of NWACS-FULL which is time consuming to update and unlikely to be useful for providing ERPs. It is potentially useful for scenario modeling, i.e., asking what-if questions related to alternative future developments where historic estimates and data are no longer valid in the future, for example, to address management objectives to minimize risks due to shifting environmental drivers.

VADER is a statistical-catch-at-age model similar to BAM. A benefit over EwE based models is that it is able to estimate F and other management-relevant quantities directly from data. It could be used for Operating Model (OM) conditioning and has an important benefit of being able to condition OM scenarios based on structural uncertainty, which could be used to evaluate the robustness of advice based on BAM.

A surplus production model with time varying r was also used, which I see more as an alternative to the single species assessment (see section on ToR for single species).

Ecosystem management objectives and corresponding performance metrics had been agreed at a 2015 stakeholder workshop. These related to i) sustainability of menhaden, ii) impact of the fishery on predators, and iii) minimisation of risk due to changing environmental conditions. All models were appropriate for i), only VADER, NWACS-MICE and NWACS-FULL can address ii) and no models are currently set up to address iii).

It appears that the NWACS-MICE is currently the best model to address ecosystem management objectives when combined with BAM. The VADER model may be useful for conditioning OMs for use in MSE, as the two models use many of the same datasets and produce similar outputs. This would make coupling them in a feedback loop possible.

When conditioning an OM, the objective is to develop hypotheses about stock dynamics that are then used to test the robustness of the methods used to provide advice. This requires OMs to be conditioned on different datasets, and potentially for OM scenarios to be weighted. For example, to evaluate the robustness of advice based on BAM to assumptions about time invariant M, an OM conditioned using VADER would allow M to vary with time. This could evaluate the robustness of assuming M to be time invariant in BAM when large variability in M was seen in the ecosystem models.

ToR 4. Evaluate the methods used to estimate reference points and total allowable catch.

The trade-offs between menhaden and bass reference points were well explained and presented. e.g., using NWACS-MICE (Figure 20).

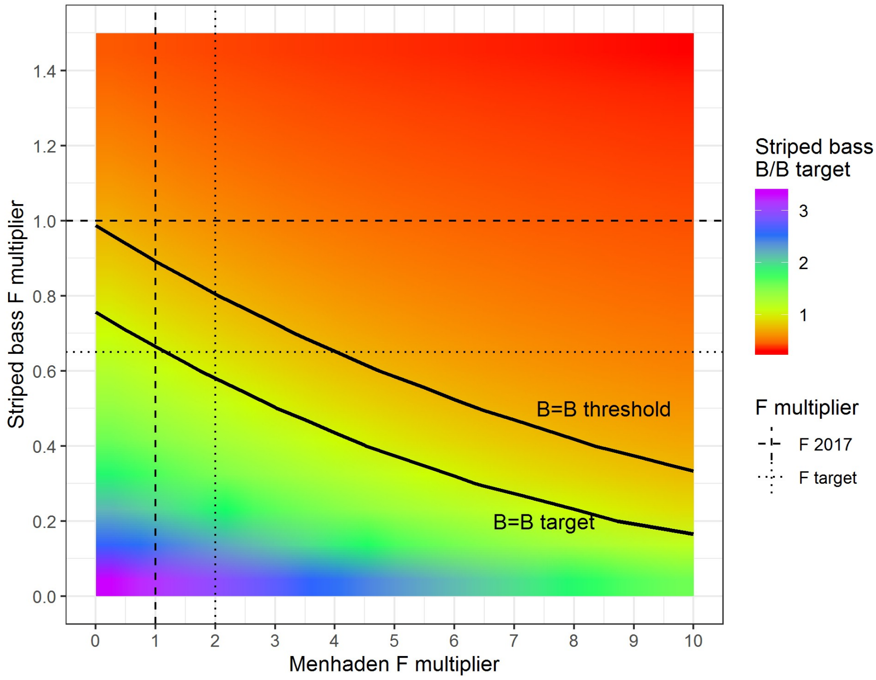

Figure 20: Striped bass age 6+ biomass ratio ($B/BTARGET$) in the terminal year of the NWACS-MICE projections as a function of fishing mortality on both Atlantic menhaden and striped bass. The solid black lines represent the contours where striped bass $B = BTHRESHOLD$ and $B = BTARGET$ The dashed lines highlight specific F scenarios where F is equivalent to the F in 2017 or the single-species F target for each species.

NWACS-MICE and BAM could be used to develop a scientific management framework, NWACS- FULL for scenario modeling and VADER to condition OMs to simulation test stock assessment models and management procedures (MPs) based on them (e.g., BAM or surplus production models). In addition, empirical MP (i.e., those where data rather than a model are used to set management action) could be evaluated.

MSE could be used to examine alternative scenarios to ensure the management advice is robust. Ideally a stepwise approach should be taken to ensure that the work does not become overwhelming. A first step is to agree on the OM. VADER and EwE-MICE are likely candidates. VADER could also be used relatively easily as an OM for use without feedback to generate inputs for BAM.

ToR 5. Evaluate the diagnostic analyses performed as appropriate to each model

The main diagnostic tool used was sensitivity analysis, where for a particular model a base case is agreed and scenarios corresponding to alternative sets of assumptions are run and the outputs compared. A problem with this approach is that a range of models with different structures and data requirements were used. Also, in some cases the models required outputs from other models as inputs. This means that the impacts of different scenarios cannot be set up or compared systematically. This is compounded by the fact that models are evaluated based on model outputs (such as harvest rate or age 1+ biomass) which cannot be observed. This means that comparisons between and within models can be somewhat subjective.

The scenarios for the Surplus production models were based on choice of CPUE series (RCPUE and PRFC) and start year, while for VADER the sensitivity runs compared the model with and without trophic interactions. This produced counter-intuitive results, possibly due to problems with the proportion of total mortality (Z) allocated to predation, which need further evaluation.

For the EwE models NWACS-MICE and NWACS-FULL, a suite of sensitivity runs were conducted with alternative dynamic (Ecosim) parameterisations. For the NWACS-FULL sensitivities explored model behavior with and without vulnerability caps, with and without manual adjustments to selected parameters, and with observed and increased diet proportions of menhaden for predators. The NWACS-MICE sensitivities explored similar parameterizations to NWACS-FULL as well as the effect of EwE-estimated primary production anomalies. A final sensitivity examined impacts of fitting to recruitment deviations as well as increasing the prey-switching exponent. Sensitivity runs for the NWACS-FULL suggested that manual tuning of parameters was necessary to balance model fits to biomass with reasonable stock-recruitment dynamics. While for the NWACS-MICE (fitted to indices rather than stock assessment outputs), sensitivity runs demonstrated that vulnerability caps reduced or eliminated model instabilities in projections. It was found during the meeting that small changes in some parameters can have significant impacts on model outputs.

Retrospective analysis was performed for VADER (only for a three-year peel) and the production models. For the production models, removal of up to four years of data from the end of the time series had little effect, most likely since there is little contrast in the CPUE data at the end of the time series. The outputs of the surplus production models were strongly influenced by the start time of the model, which could reflect changes in productivity.

ToR 6. Evaluate the methods used to characterize uncertainty in estimated parameters. Ensure that the implications of uncertainty in technical conclusions are clearly stated.

For the ERP models, less attention was given to estimation error, instead the focus was on model error. The comparisons presented in the ERP report generally suggest qualitative agreement across models (Figure 21), particularly when the models were adjusted for scaling differences. However, this alignment is not surprising given the common datasets used to inform the various models and the fact that a main model diagnostic was the comparison of models based on outputs, not their ability to predict observations.

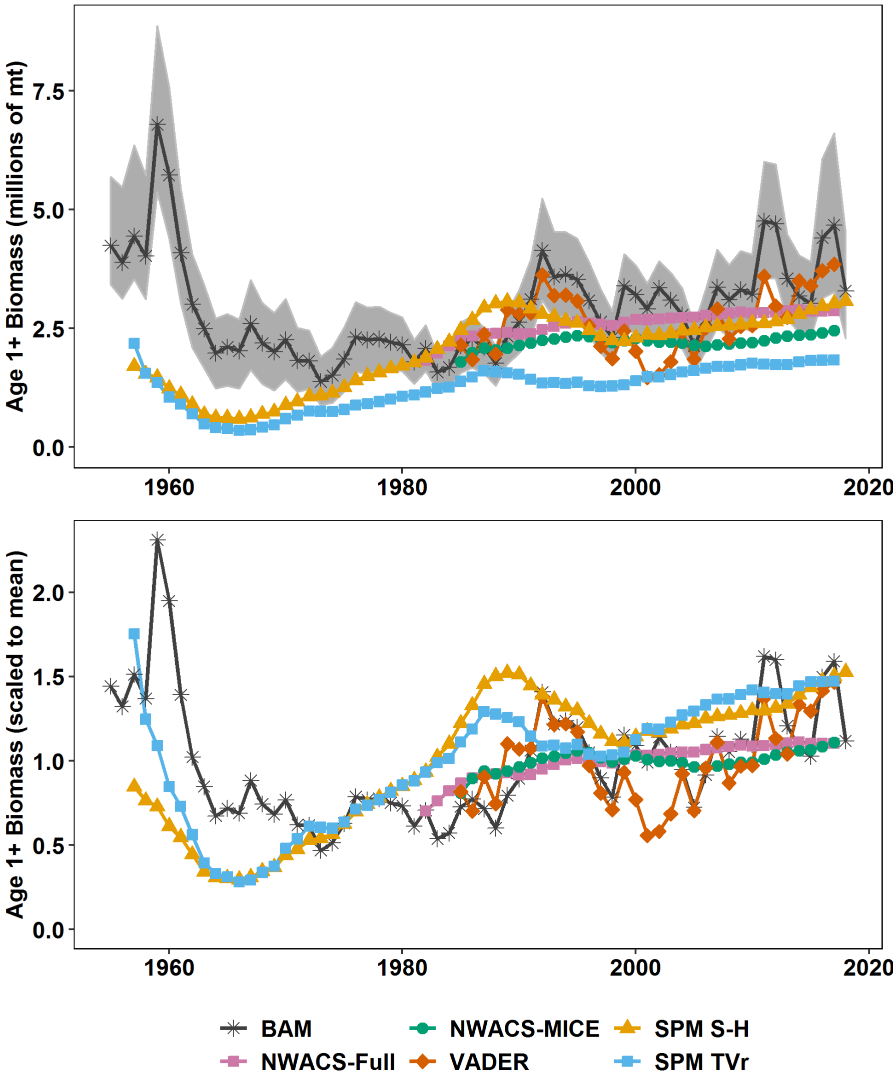

Figure 21: Comparison of ERP model runs

Importantly, when run with alternative time series which represent different fishery dynamics (e.g., the PRFC index) the models' outcomes could be quite different.

ToR 7. Minority report.

There was no minority report.

ToR 8. Recommend best estimates of stock biomass, abundance, exploitation, and stock status of Atlantic menhaden from the assessment for use in management, if possible, or specify alternative estimation methods.

Currently, the Multi-Species models are not ready to provide advice, so stock biomass, abundance, exploitation and stock status and reference points from the base run of the BAM should be used for management.

ToR 9. Review the research, data collection, and assessment methodology recommendations provided by the TC and make any additional recommendations warranted. Clearly prioritize the activities needed to inform and maintain the current assessment, and provide recommendations to improve the reliability of future assessments.

Recommendations for both short and long term research included expanding collection of data on diet and condition, non-finfish predators, and prey species. There was also a recommendation to conduct management-strategy evaluation (MSE) in order to identify harvest strategies that will meet ecosystem management objectives and to continue the development of the NWACS-MICE, NWACS-FULL and VADER models. All of which I agree with.

Due to various sources of uncertainty in the assessment of menhaden and about its role in the ecosystem, there are good reasons for conducting an MSE. This requires six steps; namely i) identification of management objectives; ii) selection of hypotheses for the OM; iii) conditioning the OM based on data and knowledge, and possible weighting and rejection of hypotheses; iv) identifying candidate management strategies; v) running the Management Procedure (MP) as a feedback control in order to simulate the long-term impact of management; and then vi) identifying the MPs that robustly meet management objectives.

A reason for the use of MSE is because the robustness of advice depends on the combination of data, estimation method, choice of reference points as well as the rules used to set management action, i.e., the Management Procedure (MP). Therefore, an Operating Model (OM) is used to represent the dynamics of the system being managed, and control actions from an MP are fed back into the OM, so that its influence on the stock and hence on future fisheries data is propagated through the stock and fishery dynamics.

The first step has been completed as the Ecosystem Management Objectives (EMO) Workshop has already identified management objectives for Atlantic menhaden. Conducting a MSE, however, can be a resource intensive process and so a multi-annual work plan, which clearly identifies the work to be done and the benefits, needs to be developed.

The next step is to agree on what are the main uncertainties about resource dynamics and how to develop an OM. The VADER model could be used since, like the BAM, it is a SSCA model and input/outputs are of the same form as the BAM that is currently used to provide advice.

The MSE could then be used to examine the robustness of the current advice without running a full feedback simulation, i.e., by using the OM to simulate alternative datasets which would help identify bias in the current assessment and the value of alternative research activities.

Conducting an MSE will be valuable in evaluating the robustness of the single species assessment to uncertainty, for example, in the parameter estimates and derived quantities (i.e., stock status relative to reference points), the assumed values of natural mortality, non-stationarity of biological processes over time, and the vulnerability of the stock to the fishery. It can therefore be used to evaluate the benefits of improved data collection and biological sampling, as well as comparing the performance of alternative harvest management strategies.

It will be important to plan an MSE process carefully, to avoid progress on management being impeded by a process that could take several years and require a large commitment of resources.

4 Summary and Recommendations

Summary

Single Species Assessment

- The single species assessment was conducted using the Beaufort Assessment Model (BAM), a Statistical Catch-at-Age Assessment (SCAA). BAM had been used in both the previous 2015 benchmark stock assessment and the 2017 stock assessment update. The BAM estimates population numbers and fishing mortality-at-age and then the population is projected forward.
- There appear to have been large changes in the life history parameters, while the indices of abundance and catch-at-age appear to have little information in them.
- The main datasets used are fishery independent indices of abundance, catch-at-age and bio- logical parameters. The fisheries dependent indices did not show strong contrast, while there appeared to be little cohort signal in the catch-at-age data, and large changes have been seen in biological parameters such as growth and fecundity.
- Fishery dependent indices in comparison showed more contrast than the fishery independent indices but the two indices developed for the ERP models were in conflict. The cause of the difference between the two fisheries derived CPUEs and the fishery independent indices could be important for management, for example, if there has been a change in the operation of the fisheries or distribution of the stock.
- A Base Case and a set of sensitivity analyses were developed to examine effects due to changes in the input data, i.e., choice of index, life history values, and ageing uncertainty. In addition, some runs addressed data choices in the ERP models.
- Uncertainty due to model structure and data choice was evaluated using sensitivity and retrospective analyses. Uncertainty in parameters (i.e., estimation error) was evaluated using a parametric Monte Carlo bootstrap (MCB) procedure in which the input data sources were re-sampled and a Markov Chain Monte Carlo (MCMC) analysis for the base which relied on the same fixed life history input as the base model run. Therefore, neither of these two approaches provides a full picture of model uncertainty. No estimates of parameter uncertainty were provided for the sensitivity runs.
- Big differences were seen in the uncertainty estimates from the MCMC and MCB. The MCMC reflects uncertainty in the ability to estimate population parameters, and hence reference points, stock status, while MCB reflects uncertainty about life history parameters and hence stock productivity. Both have implications for management advice and the ability to make forecasts.
- Sensitivity runs and retrospectives were conducted, and the main diagnostic was to compare model outcomes, i.e., estimates of SSB and F. A fuller exploration of diagnostics, e.g., runs tests and hindcasting, could have identified potential model misspecification.
- Uncertainty due to model uncertainty (i.e., structure and fixed inputs such as M) is important. This was seen in the differences across sensitivity analyses and by comparing SCCA and biomass dynamic models. A potential reason is the lack of information in the abundance and catch-at-age data.
- Likelihood profiling of key parameters such as virgin biomass or $R0$ by data components, for the Base Case could have identified the potential impacts of

the different datasets and suggested appropriate sensitivity runs to conduct or uncertainties to try and resolve.

- It will be time consuming to estimate uncertainty using MCMC and MCB, however, uncertainty could be derived from the covariance matrix. While in previous assessments it was thought that these provided an underestimate of uncertainty, they could provide a useful comparison with other methods and are quicker to run than MCB and MCMC, and so could be done for all scenarios.
- It would have been valuable to have compared estimation and model error for the sensitivity analyses. This would have shown whether the uncertainty around point estimates is greater than the uncertainty about model structure. Although it would have been difficult to do this using MCB or MCMC, the covariance of derived parameters (i.e., stock status relative to benchmarks) could have been computed from hessian. Although this approach was not used for the Base Case as it was thought to underestimate uncertainty, comparing the three forms of estimation error (MCB, MCMC and asymptotic) may have been a valuable diagnostic as it could suggest problems with the likelihood and data weightings
- The outputs of the surplus production models were strongly influenced by the start time of the model, which could reflect changes in productivity, i.e., non-stationarity. Given the large changes seen in life history parameters, the robustness of the time periods chosen to derive reference points should be evaluated, and the robustness, in particular to future conditions, evaluated.
- A major question is whether the dynamics are determined by process error or a production function. A potential approach to evaluate this is to use JABBA-Select Winker et al. (submitted) a surplus production model that incorporates life history parameters and fisheries selectivity and distinguishes between exploitable biomass and spawning biomass, enabling more direct comparison with age-structured model results.

Ecological Reference Points

- Five models were used and ranged in structural complexity and data requirements from a surplus production model with time-varying population growth rate (SPM-TVr) to a full food web model for the Northwest Atlantic continental shelf (NWACS-FULL). Intermediate models included a surplus production model that accounts for menhaden removals due to predation (Steele-Henderson), a multispecies statistical catch at age model with menhaden and five species (VADER), and a scaled down food web model (based on NWACS-FULL) that included menhaden and only a subset of key predator and prey species (NWACS-MICE).
- For uncertainty in the ERP models the main focus was on model uncertainty, i.e., conducting sensitivity analyses. However, due to the different structure of the models and different data requirements it was not possible to conduct these in a systematic way. Comparisons between models generally suggest that qualitatively the models are in agreement. However, this is not surprising given the common datasets used and the fact that a main model diagnostic was the comparison of models based on their outputs rather than their ability to predict observations.
- Some ERP models used outputs from other models, e.g., single species assessments, as inputs. This presents a possible problem for validation.

- The NWACS-MICE is currently best able to address management objectives when combined with the single species BAM assessment.
- The VADER model may be useful for conditioning Operating Models (OMs) for use in a Management Strategy Evaluation and for simulation testing BAM, and advice based upon it as the two models uses many of the same datasets.

Model Validation

- A main problem then is how to objectively compare and validate the models. Currently there is a degree of circularity, as model outputs are compared rather than how well a model can predict the data.
- A major form of model validation for both single species and ecosystem-based models was to compare sensitivity scenarios. Scenarios were compared using model outputs such as biomass and F. A reason for this is because when alternative assessment model structures are developed using different datasets, conventional model selection criteria such as AIC cannot be applied to choose between models. It is important to be able to compare model fits across sensitivity runs in an objective way. There are several approaches that have potential, e.g., runs tests (Carvalho et al., 2017) and hindcasting (Kell et al., 2016), and that could be considered in the future.
- Scientific assessment frameworks for wildlife and fisheries management are based on two main paradigms: population status and management procedure evaluation. Normally, a population dynamics model is fitted to time-series data of population indices and age-/size compositions. The model is then used to assess historical population status relative to reference points and to predict the outcomes of alternative management options. As a normal course of procedure, model diagnostics are conducted to test goodness of fit and look at residual pattern. In addition, given that development of a management procedure based on the fitted model is an ultimate goal, prediction error of the model should also be evaluated. Further, when alternative assessment model structures are developed (e.g., age-aggregated or age-structured), conventional model selection criteria such as AIC cannot be applied for evaluating those models because of the difference in the data set used. Here, we propose a hindcasting approach as a method to evaluate models through their prediction skills. Models are retrospectively re-run by removing recent years of data and the population trajectories are forecasted up to the most recent year to compare with the observed time-series of a population index (which is used in the original model fitting). We introduce examples of application of the hindcasting approach for some important global fishery resources to confirm the feasibility of the models and discuss

caveats of the approach.

Empirical Indicators

- As well as conducting analytical stock assessments, empirical indicators should be developed. These can used to look at spatial and temporal trends in stock demography, such as the relative abundance of large individuals that may make a major contribute to spawning reproductive potential. This is especially important since large changes in growth have been seen in the past and natural mortality is substantially greater than fishing mortality, see ICES Technical Guidelines for some examples[1]http://ices.dk/sites/pub/Publication Reports/Guidelines and Policies/16.04.03.02*Category3 −4ReferencePoints.pdf.*

Surplus Production

- The use of surplus production (SP) from an assessment model (Walters et al., 2008), can provide a check on whether predictions of changes in biomass ($B_{t+1}B_t$) can be made reliably based on catch and current biomass. In particular, whether similar B levels have exhibited similar SP at different historical times, i.e., whether there has been non-stationarity in production processes, i.e., are the dynamics driven by the environment or by density dependence.

Management Strategy Evaluation

- Conducting an MSE will be valuable in evaluating the robustness of the single species assessment to uncertainty, for example, in the parameter estimates and derived quantities (i.e., stock status relative to reference points), the assumed values of natural mortality, non-stationarity of biological processes over time, and the vulnerability of the stock to the fishery. It can therefore be used to evaluate the benefits of improved data collection and biological sampling as well as comparing the performance of alternative harvest strategies.
- The VADER model may be useful for conditioning Operating Models (OMs) for use in MSE, as the two models use many of the same datasets. When conditioning an OM, the objective is to develop hypotheses about stock dynamics that are then used to test the robustness of the methods used to provide advice. This requires OMs to be conditioned on different datasets, and potentially for OM scenarios to be weighted. For example, to evaluate the robustness of advice based on BAM to assumptions about time invariant M, an OM conditioned using VADER would allow M to vary with time.
- There is also potential for using an MSE to evaluate the relative value-of-information versus the value of control, i.e., whether it is better to collect more data or implement more robust management.
- However, conducting MSE will take several years and require a corresponding commitment of resources, which may result in reduced effort on other tasks. Therefore, a detailed workplan should be developed where responsibilities and potential benefits are clearly identified.

Recommendations

- If an MSE framework is going to be developed, then a multi-annual work plan needs be agreed with clear responsibilities and milestones.

- An Operating Model could be developed using either VADAR or EwE based models. Initially, since VADER is a statistical-catch-at-age model like BAM, it would be easier to use VADER without feedback to simulate data to test the robustness of BAM based advice.
- Currently, models are compared by comparing model outputs. Ideally, models should be compared by their ability to predict observations, i.e., by using cross-validation (e.g., Kell et al., 2016).
- More use should be made of model diagnostics such as residual runs tests and likelihood profiling (e.g., Carvalho et al., 2017).
- The development of empirical indicators would be useful for ecosystem monitoring. These could be simulation tested using OMs developed as part of the MSE.

5 References

F. Carvalho, A. E. Punt, Y.-J. Chang, M. N. Maunder, and K. R. Piner. Can diagnostic tests help identify model misspecification in integrated stock assessments? *Fisheries Research*, 192:28–40, 2017.

P. B. Conn. Hierarchical analysis of multiple noisy abundance indices. *Canadian Journal of Fisheries and Aquatic Sciences*, 67(1):108–120, 2009.

H. Gislason, J. Pope, J. Rice, and N. Daan. Coexistence in north sea fish communities: implications for growth and natural mortality. *ICES J. Mar. Sci.*, 65(4):514–530, 2008.

A. Jensen. Origin of relation between k and l inf and synthesis of relations among life history parameters. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(5):987–989, 1997.

L. T. Kell, P. Levontin, C. R. Davies, S. Harley, D. S. Kolody, M. N. Maunder, I. Mosqueira, G. M. Pilling, and R. Sharma. The quantification and presentation of risk. *Management Science in Fisheries: An Introduction to Simulation-Based Methods*, page 348, 2015.

L. T. Kell, A. Kimoto, and T. Kitakado. Evaluation of the prediction skill of stock assessment using hindcasting. *Fisheries Research*, 183:119–127, 2016.

R. D. Methot. Technical description of the stock synthesis II assessment program version 1.17-March 2005. *Unpublished draft report provided on the CD-ROM of background materials for the STAR. NOAA Fisheries, Seattle, Washington, USA*, 2005.

C. J. Walters, R. Hilborn, and V. Christensen. Surplus production dynamics in declining and recovering fish populations. *Canadian Journal of Fisheries and Aquatic Sciences*, 65(11):2536–2551, 2008.

H. Winker, F. Carvalho, J. Thorson, L. Kell, D. Parker, M. Kapur, R. Sharma, A. Booth, and S. Kerwath. Jabba-select: Incorporating life history and fisheries' selectivity into surplus production models. *Fisheries Research*, submitted.

A Bibliography of Review Materials

Document #	Title	Author
SEDAR 69 – SAR1	Assessment of Atlantic Menhaden Single Species Benchmark Report	To be prepared by SEDAR 69
SEDAR 69 – SAR2	Assessment of Atlantic Menhaden Ecological Reference Point Report	To be prepared by SEDAR 69
	ASMFC Instructions for Reviewers	
	Supplementary Materials	
SEDAR 69 – RD01	SEDAR 40 Stock Assessment Report Atlantic Menhaden	SEDAR 2015
SEDAR69 – RD02	Hierarchical analysis of multiple noisy abundance Indices	P. Conn 2010
SEDAR 69 – RD03	Estimation of movement and mortality of Atlantic menhaden during 1966–1969 using a Bayesian multi-state mark-recovery model	Liljestrand et.al. 2019
SEDAR 69 – RD04	Trends in Relative Abundance and Early Life Survival of Atlantic Menhaden during 1977–2013 from Long-Term Ichthyoplankton Programs	Simpson et.al. 2016
SEDAR 69 – RD05	Multi-state dead recovery mark-recovery model performance for estimating movement and mortality rates	Liljestrand et. al. 2019
SEDAR 69 – RD06	A MULTISPECIES STATISTICAL CATCH-ATAGE (MSSCAA) MODEL FOR A MIDATLANTIC SPECIES COMPLEX	McNamee, 2018
SEDAR 69 – RD07	Evaluating the performance of a multispecies statistical catch-at-age model	Curti, 2013
SEDAR 69 – RD08	Parameter estimation in Stock Assessment Modelling: Caveats with Gradient-based algorithms	Subbey, 2018
SEDAR 69 – RD09	Reconciling single-species TACs in the North Sea demersal fisheries using the Fcube mixed-fisheries advice framework	Ulrich et.al. 2011
SEDAR 69 – RD10	Working Group on Mixed Fisheries Advice (WGMIXFISH-ADVICE)	ICES Advisory Committee, 2016
SEDAR 69 – RD11	Evaluation of Current and Alternative Harvest Control Rules for Blue Whiting Management using Hindcasting	Kell and Levontin, 2019
SEDAR 69 – RD12	Public comment Forum Submissions	SEDAR, 2019
SEDAR 69 – RD13	Cookbook for Using Model Diagnostics in Integrated Stock Assessments	Carvalho, 2019

B Performance Work Statement

Center for Independent Experts (CIE) Program
External Independent Peer Review
SEDAR 69 Atlantic Menhaden Assessment Review

Background

The National Marine Fisheries Service (NMFS) is mandated by the Magnuson-Stevens Fishery Conservation and Management Act, Endangered Species Act, and Marine Mammal Protection Act to conserve, protect, and manage our nation's marine living resources based upon the best scientific information available (BSIA). NMFS science products, including scientific advice, are often controversial and may require timely scientific peer reviews that are strictly independent of all outside influences. A formal external process for independent expert reviews of the agency's scientific products and programs ensures their credibility. Therefore, external scientific peer reviews have been and continue to be essential to strengthening scientific quality assurance for fishery conservation and management actions.

Scientific peer review is defined as the organized review process where one or more qualified experts review scientific information to ensure quality and credibility. These expert(s) must conduct their peer review impartially, objectively, and without conflicts of interest. Each reviewer must also be independent from the development of the science, without influence from any position that the agency or constituent groups may have. Furthermore, the Office of Management and Budget (OMB), authorized by the Information Quality Act, requires all federal agencies to conduct peer reviews of highly influential and controversial science before dissemination, and that peer reviewers must be deemed qualified based on the OMB Peer Review Bulletin standards.
http://www.cio.noaa.gov/services_programs/pdfs/OMB_Peer_Review_Bulletin_m05-03.pdf).
Further information on the CIE program may be obtained from www.ciereviews.org.

Scope

The SouthEast Data, Assessment, and Review (SEDAR) is the cooperative process by which stock assessment projects are conducted in NMFS' Southeast Region. SEDAR was initiated to improve planning and coordination of stock assessment activities and to improve the quality and reliability of assessments.

SEDAR 69 will be a CIE assessment review conducted for ASMFC Atlantic menhaden. The review workshop provides an independent peer review of SEDAR stock assessments. The term review is applied broadly, as the review panel may request additional analyses, error corrections and sensitivity runs of the assessment models provided by the assessment panel. The review panel is ultimately responsible for ensuring that the best possible assessment is provided through the SEDAR process. The stocks assessed through SEDAR 69 are within the jurisdiction of the Atlantic States Marine Fisheries Commission and the states of Florida, Georgia, South Carolina, North Carolina, Virginia, Maryland, Delaware, Pennsylvania, New Jersey, New York, Connecticut, Rhode Island, Massachusetts, New Hampshire, and Maine. The specified format and contents of the individual peer review reports are found in Annex 1. The Terms of Reference (TORs) of the peer review are listed in Annex 2. Lastly, the tentative agenda of the panel review meeting is attached in Annex 3.

Requirements

NMFS requires three (3) reviewers to conduct an impartial and independent peer review in accordance with the Performance Work Statement (PWS), OMB guidelines, and the TORs below. The reviewers shall have a working knowledge in stock assessment, statistics, fisheries science, and marine biology sufficient to complete the primary task of providing peer-review advice in compliance with the workshop Terms of Reference fisheries stock assessment. It would be preferable for CIE reviewers to have expertise in forage fish population dynamics, Statistical Catch-at-Age modeling, Multispecies/Ecosystem Models with a focus on Multispecies Statistical Catch-at-Age models and Ecopath with Ecosim models, menhaden/forage fish life history and ecology, and/or management strategy evaluations/decisional frameworks.

Tasks for Reviewers

1) Two weeks before the peer review, the NMFS Project Contacts will send (by electronic mail or make available at an FTP site) to the CIE reviewers the necessary background information and reports for the peer review. In the case where the documents need to be mailed, the NMFS Project Contacts will consult with the contractor on where to send documents. CIE reviewers are responsible only for the pre-review documents that are delivered to the reviewer in accordance to the PWS scheduled deadlines specified herein. The CIE reviewers shall read all documents in preparation for the peer review.
2) Attend and participate in the panel review meeting. The meeting will consist of presentations by NOAA and other scientists, stock assessment authors and others to facilitate the review, to answer any questions from the reviewers, and to provide any additional information required by the reviewers.
3) After the review meeting, reviewers shall conduct an independent peer review report in accordance with the requirements specified in this PWS, OMB guidelines, and TORs, in adherence with the required formatting and content guidelines; reviewers are not required to reach a consensus.
4) Each reviewer should assist the Chair of the meeting with contributions to the summary report.
5) Deliver their reports to the Government according to the specified milestones dates.

Foreign National Security Clearance

When reviewers participate during a panel review meeting at a government facility, the NMFS Project Contact is responsible for obtaining the Foreign National Security Clearance approval for reviewers who are non-US citizens. For this reason, the reviewers shall provide requested information (e.g., first and last name, contact information, gender, birth date, passport number, country of passport, travel dates, country of citizenship, country of current residence, and home country) to the NMFS Project Contact for the purpose of their security clearance, and this information shall be submitted at least 30 days before the peer review in accordance with the NOAA Deemed Export Technology Control Program NAO 207-12 regulations available at the Deemed Exports NAO website: http://deemedexports.noaa.gov/ and http://deemedexports.noaa.gov/compliance_access_control_procedures/noaa-foreign-nationalregistration-system.html. The contractor is required to use all appropriate methods to safeguard Personally Identifiable Information (PII).

Place of Performance

The place of performance shall be at the contractor's facilities, and in Charleston, SC.

Period of Performance

The period of performance shall be from the time of award through January 2020. Each CIE reviewer's duties shall not exceed 14 days to complete all required tasks.

Schedule of Milestones and Deliverables: The contractor shall complete the tasks and deliverables in accordance with the following schedule.

Within two weeks of award	Contractor selects and confirms reviewers
2 weeks prior to the panel review	Contractor provides the pre-review documents to the reviewers
November 4-8, 2019	Panel review meeting
Approximately 3 week later	Contractor receives draft reports
Within 2 weeks of receiving draft reports	Contractor submits final reports to the Government

Applicable Performance Standards
The acceptance of the contract deliverables shall be based on three performance standards: (1) The reports shall be completed in accordance with the required formatting and content; (2) The reports shall address each TOR as specified; and (3) The reports shall be delivered as specified in the schedule of milestones and deliverables.

Travel
All travel expenses shall be reimbursable in accordance with Federal Travel Regulations (http://www.gsa.gov/portal/content/104790). International travel is authorized for this contract. Travel is not to exceed $10,000.

Restricted or Limited Use of Data
The contractors may be required to sign and adhere to a non-disclosure agreement.

Project Contacts:
Larry Massey – NMFS Project Contact
150 Du Rhu Drive, Mobile, AL 36608
(386) 561-7080
larry.massey@noaa.gov
Kathleen Howington - SEDAR Coordinator
Science and Statistics Program
South Atlantic Fishery Management Council
4055 Faber Place Drive, Suite 201
North Charleston, SC 29405
Kathleen.howington@safmc.net

Annex 1: Peer Review Report Requirements

1. The report must be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether the science reviewed is the best scientific information available.
2. The report must contain a background section, description of the individual reviewers' roles in the review activities, summary of findings for each TOR in which the weaknesses and strengths are described, and conclusions and recommendations in accordance with the TORs.
a. Reviewers must describe in their own words the review activities completed during the panel review meeting, including a brief summary of findings, of the science, conclusions, and recommendations.
b. Reviewers should discuss their independent views on each TOR even if these were consistent with those of other panelists, but especially where there were divergent views.
c. Reviewers should elaborate on any points raised in the summary report that they believe might require further clarification.
d. Reviewers shall provide a critique of the NMFS review process, including suggestions for improvements of both process and products.
e. The report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed, regardless of whether or not they read the summary report.
The report shall represent the peer review of each TOR, and shall not simply repeat the contents of the summary report.
3. The report shall include the following appendices:
Appendix 1: Bibliography of materials provided for review
Appendix 2: A copy of this Performance Work Statement
Appendix 3: Panel membership or other pertinent information from the panel review meeting.

Annex 2. Terms of reference

TERMS OF REFERENCE

For the 2019 ASMFC Atlantic Menhaden Single-Species Benchmark Peer Review and
2019 ASFMC Atlantic Menhaden Ecological Reference Points Benchmark Peer Review

Terms of Reference for the Atlantic Menhaden Single-Species Peer Review

1. Evaluate the thoroughness of data collection and the presentation and treatment of fishery-dependent and fishery-independent data in the assessment, including the following but not limited to:
    a. Presentation of data source variance (e.g., standard errors).
    b. Justification for inclusion or elimination of available data sources,
    c. Consideration of data strengths and weaknesses (e.g., temporal and spatial scale, gear selectivities, aging accuracy, sample size),
    d. Calculation and/or standardization of abundance indices.

2. Evaluate the methods and models used to estimate population parameters (e.g., F, biomass, abundance) and biological reference points, including but not limited to:
    a. Evaluate the choice and justification of the preferred model(s). Was the most appropriate model (or model averaging approach) chosen given available data and life history of the species?
    b. If multiple models were considered, evaluate the analysts' explanation of any differences in results.
    c. Evaluate model parameterization and specification (e.g., choice of CVs, effective sample sizes, likelihood weighting schemes, calculation/specification of M, stock-recruitment relationship, choice of time-varying parameters, plus group treatment).

3. Evaluate the diagnostic analyses performed, including but not limited to:
    a. Sensitivity analyses to determine model stability and potential consequences of major model assumptions
    b. Retrospective analysis

4. Evaluate the methods used to characterize uncertainty in estimated parameters. Ensure that the implications of uncertainty in technical conclusions are clearly stated.

5. If a minority report has been filed, review minority opinion and any associated analyses. If possible, make recommendation on current or future use of alternative assessment approach presented in minority report.

6. Recommend best estimates of stock biomass, abundance, and exploitation from the assessment for use in management, if possible, or specify alternative estimation methods.

7. Evaluate the choice of reference points and the methods used to estimate them. Recommend stock status determination from the assessment, or, if appropriate, specify alternative methods/measures.

8. Review the research, data collection, and assessment methodology recommendations provided by the TC and make any additional recommendations warranted. Clearly prioritize the activities needed to inform and maintain the current assessment, and provide recommendations to improve the reliability of future assessments.

9. Recommend timing of the next benchmark assessment and updates, if necessary, relative to the life history and current management of the species.

10. Prepare a peer review panel terms of reference and advisory report summarizing the panel's evaluation of the stock assessment and addressing each peer review term of reference. Develop a list of tasks to be completed following the workshop. Complete and submit the report within 4 weeks of workshop conclusion.

Terms of Reference for Atlantic Menhaden Ecological Reference Points Peer Review

1. Evaluate the justification for the inclusion, elimination, or modification of data from the Atlantic menhaden single-species benchmark assessment.

2. Evaluate the thoroughness of data collection and the presentation and treatment of additional fishery-dependent and fishery-independent data sets in the assessment, including but not limited to:
    a. Presentation of data source variance (e.g., standard errors).
    b. Justification for inclusion or elimination of available data sources,
    c. Consideration of data strengths and weaknesses (e.g., temporal and spatial scale, gear selectivities, aging accuracy, sample size),
    d. Calculation and/or standardization of abundance indices.

3. Evaluate the methods and models used to estimate Atlantic menhaden population parameters (e.g., F, biomass, abundance) that take into account Atlantic menhaden's role as a forage fish, including but not limited to:
    a. Evaluate the choice and justification of the recommended model(s). Was the most appropriate model (or model averaging approach) chosen given available data and life history of the species?
    b. If multiple models were considered, evaluate the analysts' explanation of any differences in results.
    c. Evaluate model parameterization and specification as appropriate for each model (e.g., choice of CVs, effective sample sizes, likelihood weighting schemes, calculation/specification of M, stock-recruitment relationship, choice of time-varying parameters, choice of ecological factors).

4. Evaluate the methods used to estimate reference points and total allowable catch.

5. Evaluate the diagnostic analyses performed as appropriate to each model, including but not limited to:
    d. Sensitivity analyses to determine model stability and potential consequences of major model assumptions
    e. Retrospective analysis

6. Evaluate the methods used to characterize uncertainty in estimated parameters. Ensure that the implications of uncertainty in technical conclusions are clearly stated.
7. If a minority report has been filed, review minority opinion and any associated analyses. If possible, make recommendation on current or future use of alternative assessment approach presented in minority report.

8. Recommend best estimates of stock biomass, abundance, exploitation, and stock status of Atlantic menhaden from the assessment for use in management, if possible, or specify alternative estimation methods.

9. Review the research, data collection, and assessment methodology recommendations provided by the TC and make any additional recommendations warranted. Clearly prioritize the activities needed to inform and maintain the current assessment, and provide recommendations to improve the reliability of future assessments.

10. Recommend timing of the next benchmark assessment and updates, if necessary, relative to the life history and current management of the species.

11. Prepare a peer review panel terms of reference and advisory report summarizing the panel's evaluation of the stock assessment and addressing each peer review term of reference. Develop a list of tasks to be completed following the workshop. Complete and submit the report within 4 weeks of workshop conclusion.

Annex 3. Agenda

Tentative Agenda (Draft 08.02.19)

SEDAR 69 Atlantic Menhaden & Ecological Reference Points Review Workshop

Charleston, South Carolina
November 4-8, 2019

Monday

9:00 a.m.	**Convene**	
9:00 a.m. – 9:20 a.m.	**Introductions and Opening Remarks**	
	Coordinator/Chair	
	- Agenda Review, TOR, Task Assignments	
9:20 a.m. – 11:00 a.m.	**Assessment Presentations: Atlantic menhaden**	
	- Assessment History	*Amy Schueller*
	- Life History	*Amy Schueller*
	- Regulatory History	*Max Appelman*
	- Commercial Reduction Fishery	*Ray Mroch*
	- Commercial Bait Fishery	*Kristen Anstead*
	- Indices of Abundance	*Kristen Anstead*
11:00 a.m. – 11:15 a.m.	**Break**	
11:15 a.m. – 12:15 p.m.	**Continue Assessment Presentations**	
	- Assessment Model and Results	*Amy Schueller*
12:15 p.m. – 1:30 p.m.	**Lunch Break**	
1:30 p.m. – 3:30 p.m.	**Continue Assessment Presentations**	
	- Reference Points and Stock Status	*Amy Schueller*
	- Projection Methodology	*Amy Schueller*
	- Research and Modeling Recommendations	*Kristen Anstead*
3:30 p.m. – 3:45 p.m.	**Break**	
3:45 p.m. – 4:45 p.m.	**Panel Discussion**	**Chair**
	- Begin discussion with SAS	
	- Identify additional analyses, sensitivities, corrections	
4:45 p.m. – 5:15 p.m.	**Panel Comments**	**Chair**
	- Initial panel comments on assessment	
5:15 p.m. – 5:45 p.m.	**Day 1 Summary & assignments to analytical team**	**Chair**
5:45 p.m. – 6:00 p.m.	**Public Comment**	

***Monday Goals*:** Initial single-species assessment presentations completed, sensitivity and base model discussion begun, additional analyses requested

Tuesday

8:30 a.m. – 9:00 a.m.	**Review additional single-species analyses**	*Amy Schueller*
9:00 a.m. – 10:30 a.m.	**Ecological Reference Points Assessment**	
	-Ecological Modeling Objectives	*Matt Cieri*
	-Modeling History	
	-Predator & Prey Choices	

	-*Multispecies Data*	*Katie Drew*
10:30 a.m. – 10:45 a.m.	**Break**	
10:45 a.m. – 11:45 a.m.	**Ecosystem Modeling Presentations**	
	Multispecies Surplus Production Models	*Katie Drew*
11:45 a.m. – 12:15 p.m.	**Panel Discussion**	**Chair**
	- *Discussion on surplus production models*	
	- Identify additional analyses to be requested	
12:15 p.m. – 1:30 p.m.	**Lunch Break**	
1:30 p.m. – 2:30 p.m.	**Ecosystem Modeling Presentations Continued**	
	Multispecies Statistical Catch-at-Age Model	*Jason McNamee*
2:30 p.m. – 3:15 p.m.	**Panel Discussion**	**Chair**
	-Discussion of MSSCAA model	
	-Identify additional analyses to be requested	
3:15 p.m. – 3:30 p.m.	**Break**	
3:30 p.m. – 4:30 p.m.	**Ecosystem Modeling Presentations Continued**	
	Ecopath with Ecosim Models	*Dave Chagaris*
4:30 p.m. – 5:15 p.m.	**Panel Discussion**	**Chair**
	-Discussion of EwE models	
	-Identify additional analyses to be requested	
5:15 p.m. – 5:45 p.m.	**Day 2 Summary & assignments to analytical team**	**Chair**
5:45 p.m. – 6:00 p.m.	**Public Comment**	

Tuesday Goals**:** Initial ecosystem model presentations completed, sensitivity and base model discussion begun, additional analyses requested

<u>Wednesday</u>

8:30 a.m. – 10:30 a.m.	**Ecological Reference Points Presentation**	
	- Review & Synthesis of Result	*Matt Cieri &*
	- Management & reference points recommendations	*Dave Chagaris*
10:30 a.m. – 11:00 a.m.	**Break**	
11:00 a.m. – 12:00 p.m.	**Panel Discussion**	**Chair**
	- *Ecological reference points & management*	
	- Identify additional analyses to be requested	
12:00 p.m. – 1:30 p.m.	**Lunch Break**	
1:30 p.m. – 3:30 p.m.	**Continue Panel Discussion**	**Chair**
	- *Ecological reference points & management*	
	- Identify additional analyses to be requested	
3:30 p.m. – 4:00 p.m.	**Break**	
4:00 p.m. – 5:00 p.m.	**Review additional ecosystem modeling analyses**	*TBD*
5:00 p.m. – 5:45 pm.	**Day 3 Summary & assignments to analytical team**	**Chair**
5:45 p.m. – 6:00 p.m.	**Public Comment**	

Wednesday Goals**:** Initial review and discussion of reference points and management recommendations

<u>Thursday</u>

8:30 a.m. – 10:30 a.m.	**Panel Discussion**	**Chair**
	- Final menhaden analyses & projections reviewed	
10:30 a.m. – 11:00 a.m.	**Break**	
11:00 a.m. – 12:00 p.m.	**Panel Discussion**	**Chair**
	-Single-species discussions continues	
12:00 p.m. – 1:30 p.m.	**Lunch Break**	
1:30 p.m. – 3:30 p.m.	**Panel Discussion**	**Chair**
	- Final ecosystem analyses reviewed	
3:30 p.m. – 4:00 p.m.	**Break**	

4:00 p.m. – 5:45 p.m.	**Panel Discussion** *- Ecological reference points assessment*	**Chair**
5:45 p.m. – 6:00 p.m.	**Public Comment**	
Friday		
8:30 a.m. – 10:30 a.m.	**Panel Discussion/Panel Work Session** *- Continue deliberations* *- Recommendations and comments*	**Chair**
10:30 a.m. – 11:00 a.m.	**Break**	
11:00 a.m. – 12:30 p.m.	**Panel Discussion or Work Session** *- Review Reports*	**Chair**
12:30 p.m. – 1:00 p.m.	**Public Comment**	
1:00 p.m.	**ADJOURN**	

C Panel Membership

The expertise of the panel was broad reflecting the ToR and consisted of Dr. Michael Jones (Chair), and Council of Independent Expert reviewers Dr. Kenneth Frank, Dr. Laurence Kell, and Dr. Daniel Howell. In addition, although not a CIE reviewer, Dr. Sarah Gaichas was a member of the review panel. Dr. Michael Jones is Professor Emeritus at the Quantitative Fisheries Center at Michigan State university. Dr. Kenneth Frank is Research Scientist at Fisheries and Oceans Canada. Dr. Laurence Kell is Visiting Professor in Fisheries Management at Imperial College London. Dr Daniel Howell is Research Professor at IMR, Norway. Dr. Sarah Gaichas is Research Fisheries Biologist at NOAA.