Summary Report of the Applied State Space Models Research Track Stock Assessment Peer Review

February 12-15, 2024
Northeast Fisheries Science Center, Woods Hole, Massachusetts

Report prepared by Panel Members:
Dr. Yong Chen (Chair), NEFMC SSC
Dr. Anders Nielsen, independent contractor with CIE
Dr. Noel Cadigan, independent contractor with CIE
Dr. Arni Magnusson, independent contractor with CIE

Introduction

The Northeast Region Coordinating Council (NRCC)[1] has developed an enhanced stock assessment process to improve the quality of assessments. The process involves two tracks of assessment work: 1) a management track that includes routine updates of previously approved assessment methods to support regular management actions (e.g., annual catch limits), and 2) a research track that allows comprehensive research and development of improved assessments on a stock-by-stock or topical basis. The research track assessment process allows for a more thorough review of information available and for the evaluation of different assessment approaches than would be possible in a standard stock assessment process where the results are immediately used for management advice. This Panel reviewed the Research Track Assessment for the topic of Applied State Space Models.

The work of the WG has been reviewed by the Applying State Space Models Research Track Peer Review Panel that met in person in the Northeast Fisheries Science Center, Woods Hole, MA from February 12-15, 2024. Online option via Webex is also available for other attendees who would like to attend the review remotely. The Panel included three independent scientists selected by the Center for Independent Experts (CIE): Dr. Anders Nielsen (National Institute of Aquatic Resources, Technical University of Denmark), Dr. Noel Cadigan (Fisheries and Marine Institute, Memorial University of Newfoundland, Canada), and Dr. Arni Magnusson (The Pacific Community, SPC). The Panel was chaired by Dr. Yong Chen (Stony Brook University and member of the Scientific and Statistical Committee of the New England Fisheries Management Council).

The Working Group Assessment Report and 12 supporting Working Papers were made available to the Panel on the data portal (https://apps-nefsc.fisheries.noaa.gov/saw/sasi/sasi_report_options.php) on January 29, 2024. The Panel was

[1] Atlantic States Marine Fisheries Commission (ASFMC), Greater Atlantic Regional Fisheries Office (GARFO), Mid-Atlantic Fishery Management Council (MAFMC), New England Fishery Management Council (NEFMC), and Northeast Fisheries Science Center (NEFSC).

also given access to the GitHub repositories used by the WG where they could access model code, data input files, and model outputs including figures and tables. Individual Panel Members and the Chair took the lead in providing first drafts of various sections of the report, but the entire Panel is responsible for the whole report. Prior to the meeting, members of the Panel met with Michele Traver (NEFSC's Stock Assessment Workshop Process Lead), Kristan Blackhart (Chief, NEFSC Population Dynamics Branch) and Alexander Dunn (Communications Specialist, NEFSC Population Dynamics Branch) to review and discuss the meeting agenda, reporting requirements, meeting logistics and the overall process.

Presentations made by WG members during the review are listed in the agenda (Appendix 2) and available as PDFs on the data portal. Other WG members were present and answered questions from the review panel and contributed to the discussions on various topics. Emily Liljestrand, Jessica Blaylock, Kiersten Curti, Dan Hennen, Tony Wood, Chris Legault, and Amanda Hart acted as rapporteurs throughout the meeting (see Appendix 4 for materials provided and Appendix 5 for meeting attendees). The WG was chaired by Tim Miller (NEFSC) and included staff from NOAA Fisheries, academia, a non-governmental organization, and state fishery management agencies. Terms of Reference for the WG are provided in Appendix 1.

Panel members and the Chair drafted this Summary Report in a Google Doc. The Panel Chair compiled and edited this Summary Report with assistance (by correspondence) from the CIE Panelists, before submission of a draft report to the WG. The scope of the WG review of the draft was limited to suggesting corrections for errors of fact or requesting that Panel recommendations be clarified. Additionally, each of the CIE Panelists will submit their separate reviewer's reports to the CIE.

The Panel concluded that TORs 1-3 and 5 ***were fully met*** and TOR 4 ***was not met***. The Panel agrees that the state-space model Woods Hole Assessment Model (WHAM) is a significant advancement from the traditional models in providing a formal modeling framework to explicitly account for time-varying biological and fishery parameters and ecosystem/environmental covariates in stock assessment. Based on the literature review and extensive computer simulation work conducted in this research track assessment, the WG developed the guidelines to inform diagnosing and selecting preferred state space model configurations and to explicitly incorporate ecosystem and environmental effects in assessment models. The capacity to explicitly model various process errors in life history and fishery processes is especially important for fisheries stock assessments in the northeast USA, given the rapid changing ecosystem in the region. The Panel agreed that the WHAM model selected using the guidelines developed in this study is likely to perform better than traditional methods in providing management advice, including estimating biological reference points (BRPs) and making projections. The Panel recommends that the Center provide resources to complete the close-loop or MSE type of simulation work to compare relative performance of tradition and state-space models regarding management metrics (i.e., TOR 4). The Panel agreed that the 4 case studies conducted in this research track showed that the WHAM can improve the quality of stock assessment and can be considered and further developed for the use in management track assessments. However, the Panel cautioned that the 4 case studies included in this research track were mainly done to showcase the utility and advantages of the WHAM and the results derived in these case studies are not intended to inform management.

The Panel's evaluation of the WG's response to the five TORs is provided below and concludes with a summary of key recommendations.

Evaluation of the Terms of Reference for Applied State Space Models

TOR 1. Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.

The Panel concluded that this TOR had been **fully met**.

The WG addressed this ToR through review of: 1) the scientific literature on state-space modeling, 2) the scientific literature on state-space stock assessment modeling, 3) the scientific literature on assessment model diagnostics, and 4) of relevant results from working papers of extensive simulation studies prepared by members of the WG. These reviews and simulation studies formed the basis of recommended practices for selecting among alternative configurations of state-space stock assessment models.

The WG considered that there are several qualities that can be assessed to determine a preferred assessment model structure. These qualities include but are not limited to:

1. Better representation of realism of the biology and data generating mechanisms. The WG felt that the primary way that state-space models (SSMs) may improve biological realism is by accounting for and estimating temporal variability in the demographic parameters that are otherwise treated as constant over time as well as the variance and autocorrelation of these processes. These parameters include recruitment, growth, and survival/natural mortality, but also fisheries and possibly survey selectivity, and survey catchability.
2. Statistical reliability. The Panel suggests that this is an important advantage of state-space stock assessment models that more effectively model high-dimensional parameters (e.g., time-varying parameters) as random effects. SSMs separately account for process and observation errors. These models are directly useful for stochastic projections. An important and novel development in WHAM is the structural errors in variables (SEV) methodology to address covariate measurement errors. WHAM provides the ability to link covariates to several stock productivity processes (see TOR 3) and the Panel suggests that the SEV methods are widely used in statistical sciences to address problems (i.e., main bias in effect estimation) because of covariate measurement errors. The Panel felt that WHAM includes most of best-practice methodologies to account for sampling uncertainty; that is, the observational SSM likelihoods.
3. Better prediction skill. The Panel suggests that SSMs with autocorrelated process errors tend to provide improved prediction skill compared to models that do not include time-

varying productivity. The WG demonstrated from simulation analyses that missing an important source of process error can produce biased assessment results.

4. Lack of evidence of model mis-specification.

The Panel suggests that an important consideration for selecting preferred state-space model formulations is convergence. Models that do not converge frequently in simulations or retrospective analyses are not preferable for the specific stock being investigated. Alternative and usually simpler model formulations should be investigated with a good convergence rate (i.e., > 90% in simulation).

Recommendations by the WG are:

1. Treat recruitment as random effects so that variance and correlation parameters can be estimated.
2. Consider as many sources of process error as might be plausible and practical, but be aware of unintended implications for management reference points and catch advice.
3. When non-negligible mis-reporting of catch is plausible, estimation of catch process errors should be considered.
4. When reliable external estimates of observation error are available, treat them as known in the assessment model.
5. Perform posterior check of all random effects.
6. When using MASE with time-series cross-validation, the Panel recommends using the denominator as described by Hyndman and Koehler (2006). A generalization of MASE using (randomized) quantile prediction errors is needed.
7. Use a broad suite of metrics and diagnostic tools to evaluate relative performance of alternative models. Statistical reliability and AIC as a model selection tool are better when there is contrast in fishing pressure, stock size and process errors over time and more precise index and age composition observations are available.

The Panel discussed with respect to Rec. #4 whether reliable estimates of observation error for survey indices and age compositions are available. This has been a motivation to investigate model-based approaches to index standardization (e.g. Thorson et al., 2015). Variance parameters are often more difficult to estimate than mean parameters, and under-estimation may be expected with highly stratified surveys with many strata and small per stratum sample sizes.

The Panel further recommends that:

1. Recruitment process errors should normally be statistically independent of cohort process errors at older ages. This should be the default setting.
2. Estimation of M will often be difficult unless there is large contrast in F (Clark, 2022) and especially periods with low catches and F's so that most of the total mortality rates implied by survey age compositions can be attributed to M. Estimating time-variation in

M will often be more feasible (e.g., Aanes et al., 2007; Aldrin et al., 2021), but even then convergence problems have been reported in literature studies of state-space models (Cadigan, 2015; Aldrin et al., 2020)

3. The Panel agrees that some biases in parameter estimation should be expected in simulation self-tests because of the nonlinearity and data limitations. Maximum likelihood estimation is only asymptotically unbiased, and it is well known that variance parameters are often under-estimated in case studies.
4. Some smoothing bias is expected in assessment results based on simulation self-tests; however, consistent biases for a period of years are not expected. Some of the bias patterns produced for the case studies (i.e.,TOR 5) did not make sense for the Panel. This needs further research to investigate model configurations that reduce the bias.
5. Accurate estimation and partitioning of observation and process error variances is improved when there are multiple indices with common patterns in residuals that process errors can account for. Process errors have effects that are common among surveys, whereas observations errors will be unique to each survey.
6. AIC was demonstrated to be useful in model selection in some situations.

The efficacy of OSA residuals for detecting sources of model misspecification requires further research. The Panel recommends a broad suite of diagnostics should be examined, including plots of observed and predicted indices and age compositions, and retrospective analyses.

Convergence may be improved by improving starting values (i.e., internal q's, simple ASAP's for starting values, estimation in phases) plus using reduced models for cases when parameter estimates hit boundary constraints.

References

Aanes, S., Engen, S., Sæther, B.E. and Aanes, R., 2007. Estimation of the parameters of fish stock dynamics from catch-at-age data and indices of abundance: can natural and fishing mortality be separated?. Canadian Journal of Fisheries and Aquatic Sciences, 64(8), pp.1130-1142.

Aldrin, M., Tvete, I.F., Aanes, S. and Subbey, S., 2020. The specification of the data model part in the SAM model matters. Fisheries Research, 229, p.105585.

Aldrin, M., Aanes, F.L., Tvete, I.F., Aanes, S. and Subbey, S., 2021. Caveats with estimating natural mortality rates in stock assessment models using age aggregated catch data and abundance indices. Fisheries Research, 243, p.106071.

Cadigan, N.G., 2015. A state-space stock assessment model for northern cod, including under-reported catches and variable natural mortality rates. Canadian Journal of Fisheries and Aquatic Sciences, 73(2), pp.296-308.

Clark, W.G., 2022. Why natural mortality is estimable, in theory if not in practice, in a data-rich stock assessment. Fisheries Research 248, 106203.

TOR 2. Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.

The Panel concluded that this TOR had been fully met.

The research track investigated the efficacy of estimating stock-recruitment functions within state-space models and their utility in generating scientific advice via a review of current practice, first principle reasoning, and a large-scale simulation study. The computation time for the simulations related to this part easily exceeds half a year (if each model run takes about a minute). In practice these computations are done in parallel on clusters/multi-core computers to be able to present the results in time for this review.

The investigations conducted during the research track can be divided into two parts. Fitting a standard stock recruitment (e.g., Beverton-Holt) relationship and fitting a stock-recruitment relationship with environmental covariates.

Since neither stock size (SSB) nor recruitment (R) are directly observable quantities, but estimated from indirect observations (partly shared), both quantities have estimation noise and observation noise has a non-trivial covariance structure. Accounting for the covariance structure within and among the dependent (R) and the independent (SSB) variables is necessary in order to correctly estimate the stock recruitment relationship external to the state-space assessment model. Selecting and estimating the stock recruitment relationship inside the state-space assessment is the simplest way to correctly account for the covariance structure. Having the stock-recruitment function inside the state-space stock assessment model can further be useful to ensure consistency between the assessment model and short-term forecast procedures and to correctly propagate estimation uncertainty to all estimated quantities of interest.

The simulations were designed to mimic a 'typical gadid' stock in terms of number of years and number of age classes, survey fleets, growth, and selection pattern. The simulations showed - quite realistically - that estimation of a stock recruitment relationship is fragile. Successful estimation depends primarily on two things. First, a wide range of stock sizes needs to be observed and it should represent both the density dependent and density independent part of the stock recruitment relationship. The range in stock sizes were introduced in the simulations by introducing different fishing histories (constant at Fmsy or a ramp), random effects in survival or natural mortality, and/or random effects on recruitment. Secondly, the uncertainty of the recruitments around the relationship (sigmaR) should be small. If these conditions were not present it led to low convergence rates, bias in stock recruitment parameters, and poor ability to identify the correct stock recruitment relationship via AIC.

A main part of the problem regarding estimating a stock recruitment relationship is that recruitment is potentially controlled by many other factors than stock size, which leads to large noise around a pure stock recruitment relationship (sigmaR). WHAM has the unique ability to include environmental covariates in the stock recruitment relationship. This could potentially help the estimation (as the residual noise could be reduced), and it could lead to more useful models of the recruitment relationship for the purpose of forecasting recruitment when the environmental conditions are changing. A simulation study of the ability to use such

environmental covariates in the stock recruitment relationship was conducted by simulating stationary environmental processes and then having those processes drive the simulated stock recruitment processes (in three different configurations). Estimation models with correct and incorrect covariate configuration, and with constant mean recruitment were then compared. The study found a poor ability to identify the correct model via standard model selection (i.e., AIC and retrospective analysis), but also that the assessment results were relatively unaffected. The forecasts were also found to be relatively similar, which could be due to the stationarity of the simulated processes. It was found that a wide range of stock-sizes in the simulation and small uncertainty around the stock recruitment relationship reduced parameter bias and increased the ability to identify the correct model structure.

1. The research track produced the following recommendations regarding including stock recruitment relationship in state-space assessment models:

2. Consider the level of information in the stock assessment data for the stock-recruit relationship. Positive responses to these questions increase the likelihood for reliable inferences:
    a. Is the time series sufficiently long?
    b. Is there evidence of good contrast in spawning stock biomass over time?
    c. Are index and age composition observations relatively precise?
    d. Is variation in recruitment residuals (sigma-R) relatively low?

3. Estimate the stock-recruit relationship simultaneously and internal to the state-space stock assessment model.

4. Self-tests as described in TOR 1 would be prudent to confirm reliability of stock-recruit parameter estimates and biological reference points derived from them.

5. Consider alternative autocorrelation models for recruitment residuals. This will be important primarily in defining how recruitment is predicted in short-term projections.

The Panel supported these recommendations and found them well supported by the literature, the research conducted, and by reasoning from first principles. The Panel have the following comments:

1. A further good indication of the utility of including a stock recruitment relationship is if an explorative model run without a stock recruitment relationship (e.g., a constant mean or a plain random-walk recruitment model) indicate that a stock recruitment relationship is present, this can, for example, be identified by plotting the corresponding pairs of stock and recruitment.

2. In many cases, estimating a stock-recruitment curve can be problematic and should not be estimated. If the data seem informative about the shape of a stock-recruitment curve, internal estimation is well reasoned, but the study also revealed that this estimation is very fragile. Details w.r.t. the assumed distribution can be highly influential. The default assumed correlation coupling between recruitment increments and cohort increments

should be removed (or demonstrated to be warranted or at a minimum be demonstrated to be unproblematic).

3. The study clearly shows that self-tests are necessary. An additional test that should also be run is a jitter analysis, which shows that for the particular (not perfectly simulated) data set that the model parameter estimates are unique.

4. In addition to considering the correlation structure, it could also be relevant to consider the type of distribution assumed. In most cases it is likely that a standard log-normal distribution is sufficient to describe the recruitment deviations, but some stocks may have extreme recruitment events, which could better be described by another more heavy-tailed distribution.

The prospect of including environmental covariates may be better studied, and more relevant, in situations the environmental covariate is not stationary, so further research in this scenario and how to use that to inform management is warranted.

TOR 3. Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.

The Panel concluded that this TOR had been **fully met**.

The WG addressed this TOR through reviewing the best available science on including ecosystem and environmental effects in stock assessment models and conducting and analyzing an extensive and well designed operating-estimation modeling simulation studies. The review and simulation analyses formed the basis of developing guidelines for including ecosystem and environmental effects in assessment models and for estimating biological reference points and scientific advice.

The WG considered several measures that can be assessed to evaluate and include ecosystem and environmental effects in assessment models and to generate biological reference points and scientific advice. These measures include but are not limited to: convergence of the estimating models; model identifiability of an underlying relationship between environmental covariate and life history and fishery process; ΔAIC and model probability; assessment errors for fisheries parameters (e.g., recruitment, spawning stock biomass, and Fbar); biases of estimated parameters; Mohn's ρ for retrospective patterns; and projection performance relative to assumptions about the environmental covariate.

The Panel supported the following guidelines developing by the WG to include ecosystem and environmental effects in stock assessment models:

1. Limit investigations to covariates that current biological understanding suggests close links of the covariate to the particular demographic parameter.
2. Evaluate effects of covariates against models that have temporal variation in the parameter which the covariate is hypothesized to affect.
3. Check whether observation error in environmental covariates observation is low relative to other data sources as this improves reliability of inference and estimability.
4. Fix parameters describing environmental process variability where information is known.
5. Avoid the 'masking' functional form when relating stock-recruitment relationships to an environmental covariate (until further work can diagnose issues).
6. Ensure good contrast in the environmental covariate(s).
7. Conduct retrospective comparisons of models with and without covariate effects to confirm inferences are consistent as the number of years with observations changes.
8. Conduct self-tests as described in TOR 1 to confirm reliability of the estimation of effect size the covariate has on assessment model parameter estimates and reliability of biological reference points.

The Panel made the following additional comments:

(1) The current simulation design does not include scenarios where trended time-varying changes in ecosystem/environmental conditions. However, the life history (e.g., growth and natural mortality) and fishery (e.g., catchability) parameters are subject to trend changes in the northeast US. Future simulations may consider environmental covariate models with stronger effects, trends, or AR(2) dynamics. The capacity and performance of WHAM to incorporate trended process errors need to be developed and evaluated.

(2) The incorporation of environmental covariates in WHAM-based stock assessment allows for better quantification of uncertainty associated with both SSB/F and biological reference points, which can better inform scientific uncertainty in developing catch advice. However, more studies such as close-loop or MSE type of simulations (e.g., task defined in TOR 4) need to be done to compare to the traditional approach and better understand potential management implications.

(3) Ecov processes tend to affect multiple stock parameters, and multiple Ecov processes can also affect recruitment (simultaneous or sequential). This has not been considered in the current evaluation.

(4) Ecov processes can have large impacts on the projection and biological reference point estimates. This may have significant implications for developing rebuilding plans for species currently overfished. A closed-loop or MSE simulation (i.e., TOR 4) is needed to better understand management implications of using WHAM stock assessments.

(5) The study highlights the importance of data (e.g., environmental, fishery and survey) quality (i.e., low observation errors and large contrast) and quantity (i.e., data sources).

TOR 4. Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.

The Panel concluded that this TOR **had not been met**. However, the WG has drafted the following study design to complete this TOR:

A suite of operating models should be configured analogous to those done for simulation studies described by Britten et al. (WP 1) and Miller et al. (WP 5) with a groundfish life history type, a fishing fleet, and spring and fall indices representing NEFSC bottom trawl surveys. The suite of state-space operating models should span a large number of years (e.g, 100), and have alternative assumptions about process errors on recruitment, survival, and natural mortality.

The factors that should vary across operating models are

- the magnitude of observation errors (low to high),
- magnitude of variance and autocorrelation in the process errors (low to high),
- fishing history (e.g., light, moderate, and heavy historical exploitation), and
- alternative environmental covariate effects (none, small, large) or different random process errors (e.g., random walk, AR1, AR2) on recruitment, natural mortality, and/or catchability.

Using a set of seeds unique to each operating model, simulate stochastic processes and observations over some historical period. Most assessments in the region have at least 40 years of data included in the fitting, so a reasonable historical period would be at least 40 years. The fishing history during the historical period could be stock-specific, or more generic, representing different patterns in the region (see Legault et al. 2023). Starting at year 40 of the operating model, alternative WHAM estimation models would be fit to simulated observations up to year 40. The alternative WHAM models should include

- alternative state-space model configurations (e.g., alternative process error assumptions),
- alternative assumptions about environmental effects on recruitment and natural mortality, and
- a statistical catch age age model without random effects, mimicking a traditional statistical catch at age model.

Assessments in the region are typically done every 2-3 years, with catch advice based on projections over the interval between assessments. Given the focus on understanding the impacts of the state-space model on management advice, the simulations should include a single harvest control rule to reduce the potential for confounding effects of different control rules. The current acceptable biological catch (ABC) control rule used in New England applies a target F of 75% of F40% for most stocks. So, assuming three years between assessments:

- Conduct projections starting in year 41 through year 43 to determine the catch advice based on fishing at 75% of F40% (alternative projection types may be considered given the flexibility in WHAM options),

- Set catch in the operating model for years 41 to 43 with corresponding annual F determined internally (i.e., no management uncertainty),
- Re-simulate the processes and observations given the assigned random seed.

Conduct an assessment in year 43, and repeat these steps for each assessment cycle up to the end of the time frame of the operating model (e.g., 100 years).

The simulation studies completed by the working group (e.g., WPs 1–5) found poor convergence of some estimation models and the same issue could arise in this closed-loop study. A step could be added to the management model to attempt a less complex state-space model in such simulations to mimic the likely real-world strategy.

Given the completed simulations, a range of performance metrics would be calculated that summarize the state of the fishery and the stock. Such metrics could include the average catch, interannual variation in catch, average stock biomass, proportion of time the stock is overfished (both based on the true stock size and perceived by the estimation model), and the proportion of years when overfishing occurs. Comparison of performance metrics would then be made across the different operating model factors for and among each alternative WHAM assessment model to quantify management performance of each assessment model, as well as the sensitivity and tradeoffs among metrics.

The Panel supported this study design, with the following additions. To compare traditional and state-space models, the study should probably compare the estimation performance of ASAP-like WHAM vs. state-space WHAM. To address the life history factor specified in the TOR, the study design should consider life history types that are different from the generic groundfish.

TOR 5. Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

The Panel concluded that this ToR had been fully **met**.

The WG undertook bridge model runs for four groundfish stocks from the current assessment modeling platform to WHAM with an application of random effects as a potential model for management track in which further and more in-depth exploration of process errors can be conducted based on guidance and recommendations developed by this research track WG. The four stocks include Gulf of Maine haddock, Acadian redfish, Georges Bank winter founder, and Northwest Atlantic mackerel. The Panel considered that WHAM models explored for the above four stocks provide a better stock assessment framework to estimate time-varying processes and better representation of uncertainty in assessment model output. However, the four case studies only explored a limited number of configurations for the WHAM models and should not be used to discuss management implications or provide catch advice. More extensive and in-depth work is needed to understand self-test results, OSA residual distributions, and time-varying life history and fishing processes (e.g., growth, natural mortality, catchability, selectivity, and environmental covariances). The Panel's comments on each fish stock are provided below:

GOM Haddock case study:

The assessment history for GOM haddock started from comparing average exploitation rates to reference points derived from a biomass model. In 2008 the assessment was moved to an age-based ADAPT model, and in 2014 to the Age-Structured Assessment Program (ASAP), which is developed at the Wood Hole lab. The ASAP-based assessment was most recently validated at a benchmark in 2022. Retrospective patterns have been problematic in the recent history of this assessment, which was one of the reasons the review panel in 2022 suggested setting up a WHAM-based assessment for this stock.

Two age-specific surveys time series (spring and fall bottom trawls) and a combined (commercial and recreational) age-specific catch time series are available for this assessment. The raw observations consistently show occasional strong cohorts, which is common for haddock stocks, and supportive of the use of an age-based assessment model (compared to a simple biomass approach). Occasional strong cohorts are also helpful in fitting time-varying processes as done in state-space models.

The currently applied ASAP model assumes that the age-compositions from the combined catches, and the spring and fall surveys, follow a multinomial distribution. Total yearly catches from the combined catch fleet and the spring and fall surveys are assumed to follow independent log-normal distributions. Log-normal prior distributions are assumed for yearly F-deviations and for recruitment deviations. Observation variances and effective sample sizes are generally fixed (not estimated), and fishery selection blocks are pre-defined.

Four WHAM configurations were considered: An “ASAP-like”, and 3 configurations using a 2D-AR1 structure for the numbers-at-age matrix, but using 3 different observational likelihoods for the age-composition observations: multinomial, Dirichlet, and logistic normal. These configurations are compared (and partly compared to ASAP) by considering: convergence, residuals, AIC, retrospective pattern, prediction skill, and estimation performance with self-tests.

The details of the “ASAP-like” WHAM configuration were not provided in the WP, but it was explained that most details regarding fixed variances and fixed sample sizes were the same in all WHAM configurations as in the ASAP configuration (with some detailed differences for restricting initial year stock numbers-at-age). There are substantial differences in the results of the ASAP and the “ASAP-like” WHAM model configurations - especially in the last year. The “ASAP-like” WHAM model is not able to match the last year’s index as well as the ASAP. The subsequent WHAM models, which should be expected to be more conservative, were able to match the last year’s index. It seems like the last year’s index observations are ignored in the “ASAP-like” WHAM model, but the WG investigated this and found that not to be the case. The

"ASAP-like" WHAM model is not like the ASAP in this particular aspect and also not like the following WHAM configurations. This would be interesting to understand better.

One detail that is common in the three WHAM configurations with random effects on the N-at-age process is that the same correlation is assumed between the survival of neighboring age classes and between the recruitment process and survival. The Panel made a recommendation in TOR 1 about not doing this by default.

The four models are compared by the different diagnostics outlined in TOR1. The OSA residual patterns are mostly similar and all have uneven variance patterns across age groups. Retrospective patterns were less problematic for the all WHAM configurations (even for the "ASAP-like"?) compared to the ASAP model. The lowest retrospective pattern was for the multinomial observational likelihood. The prediction MASE scores for the WHAM configurations gives a mixed picture, but plots are in favor of multinomial and Dirichlet observation likelihoods. Again it appears that the last year is somehow "off" in the "ASAP-like" WHAM configuration.

The self-simulation test is clearly in favor of the "ASAP-like" WHAM configuration, which is essentially unbiased, but biases ranged between 10% and 39% for all other WHAM configurations. A "smoothing" bias is somewhat expected for state-space models, but a shift of the overall bias levels of the time series is unexpected.

The research track concluded that the WHAM configuration with the Dirichlet observational likelihood is the best of the WHAM configurations, because it is better able to match especially the last year of the survey indices, it had lower retrospective pattern, the predictions had the correct directions, and of the three WHAM with random effects, it had the lowest self-simulation bias.

This case successfully demonstrated some possible effects on the assessment results (and thereby on stock status and scientific advice) of switching to the WHAM state-space approach. The overall result of the assessment (the estimates of SSB and F time series) were substantially different and would have led to different management. The case study highlighted some issues that could be investigated further:

- The difference between the ASAP and the "ASAP-like" WHAM model is important, as that configuration is the starting point for all the further investigations. One way to approach this is to compare each likelihood part and see where the differences occur.
- The assumption of the same correlation between recruitment and survival and between survival of neighboring age classes. It could either be demonstrated that this assumption is harmless, it could be demonstrated that is valid (e.g. by producing "single joint

sample" residual for the process (Thygesen et. al 2017)), or it could be replaced with a model where the correlation is only used for the survival-part.

- The composition residuals were problematic for all model configurations. A suggestion is to use a logistic normal with a more flexible covariance structure (instead of i.i.d.).
- The bias seen between the assumed true processes and the estimated processes when simulating unbiased observations according to the models assumptions. This could be investigated by comparing with other state-space formulations (e.g., turning different parts on and off to see what triggers this).

References:

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K. and Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics, 24, pp.317-339.

Acadian redfish case study

This stock has been assessed with age-structured models since 2002. In 2008 during GARM III, an application of the Age-Structured Assessment Program (ASAP) was accepted for the assessment of this stock. This model incorporated information on the age composition of the landings, size and age composition of the population, and trends in relative abundance derived from research vessel survey biomass indices. The most recent update to the 2008 ASAP model occurred in 2023, using data through 2022. The status of the stock was determined to be not overfished and overfishing was not occurring. The retrospective pattern was classified as minor, meaning that no retrospective adjustments were required for stock status determination and short-term projections.

A long time-series of total fishery removals of Acadian redfish during 1913-2022 are used in the assessment for this stock. However, commercial age-sampling information is only available during 1969-1985 and since 2017. The ASAP model is also estimated using fall and spring bottom-trawl survey (BTS) indices. There are annual age composition estimates from the fall survey since 1975 but this information is not available for the spring survey during 1980-1984 and 1991-2016. Age compositions were available for ages 1-26+, where 26+ is a plus group.

Model diagnostics have generally been considered good; however, there is some lack of fit to surveys at the end of the assessment time series. The 2023 management track assessment peer review panel recommended that the Acadian redfish assessment be transitioned from ASAP to WHAM, which would provide greater flexibility to improve model fit to the survey indices.

This transition was explored during the review. Six WHAM configurations were presented, with the short-term goal of replicating the 2023 management track assessment ASAP model results, and the long-term goal of improving the model fit to the survey indices in a future management track or research track assessment.

Model 1. This was configured as similarly as possible to the 2023 management track assessment ASAP model. Numbers-at-age in the first year were fixed at their estimated values from the 2023 ASAP model. Recruitment was estimated using a Beverton-Holt model that included i.i.d. random effects. The steepness and recruitment scalar parameters were fixed at their estimated values from the 2023 ASAP model. Sigma R was fixed at a value of 0.8, similar to the ASAP model. Fishery selectivity and the spring and fall BTS selectivities were modeled using age-specific parameters, which were fixed at their estimated values from the 2023 ASAP model. Results demonstrated that this model fit the total catch, fall BTS index, and spring BTS index similarly compared to the 2023 ASAP model.

Model 2: Estimated logistic selectivity functions for the fishery and surveys, rather than fixing the age-specific selectivity parameters in Model 1. These models produced very similar assessment model results. OSA residuals indicated similar fits for Models 1 and 2. However, fishery age composition residuals were larger at ages 1-5 in Model 2 compared to Model 1. Model 2 had a lower AIC score than Model 1, but Model 2 had higher Mohn's rho values for F, SSB, and R compared to Model 1. The most notable difference is in the 2020 age-1 recruitment estimate, where Model 2 estimates higher recruitment than Model 1.

Model 3 estimated equilibrium numbers-at-age (NAA) in the first year rather than fixing the first year NAA parameters like in Model 2. Equilibrium recruitment (R0) is estimated while the equilibrium F is fixed near 0 (i.e., assuming an unfished stock), which seemed appropriate given the low levels of catch at the start of the assessment time series. OSA residuals indicated similar fits for Models 2 and 3. These models produced practically identical assessment model estimates, although it makes sense that Model 3 produced much wider confidence intervals for SSB in the initial model years. However, like Model 2, the Model 3 OSA age composition residuals were mostly positive, which is not expected and the Panel did not understand.

Model 4 differed from Models 1-3 in that it estimated steepness in the Beverton-Holt stock-recruitment model. OSA residuals indicated similar fits for Models 3 and 4. These models produced practically identical assessment model results. Model 4 had a comparable AIC score to Model 3. Model 4 had a higher Mohn's rho value for R, and similar Mohn's rho values for F and SSB compared to Model 3. Similar to Models 1-3, Model 4 produced mostly positive age composition OSA residuals.

Model 5 estimated sigma R to be 1.4, compared to the fixed sigma R of 0.8 in Model 4. Model 5 had a lower AIC score compared to Model 4. Model 5 had higher Mohn's rho values for F, SSB, and R compared to Model 4. OSA residuals indicated similar fits for Models 4 and 5. These models produced similar assessment results, except that Model 5 estimated higher SSB and lower F's prior to around 1960.

Model 6 did not use the BH stock-recruitment function but simply estimated a mean recruitment for the entire time-series plus i.i.d. recruitment deviations for each year. R0 was fixed at the estimated value for year-1 recruitment from the 2023 ASAP model, and sigma R was fixed at 0.8. Model 6 failed to meet the first- and second-order convergence criteria, and was not considered further.

Model 5 was proposed as the WHAM bridge run for Acadian redfish. Model 5 had the lowest AIC score of the five models, but also had the highest Mohn's rho values for F, SSB, and R. The higher Mohn's rho values were likely because Model 5 is estimating more parameters than Models 1-4. In conjunction with the improved AIC score, reducing the number of constraints on the model parameters was seen by the working group as a positive attribute of Model 5, even if it resulted in a slightly increased retrospective pattern. Simulation self-tests indicate little estimation bias. Mean percent errors were -0.4 % for F, 8.6% for R, and 1.4% for SSB. Model 5 captured the true values of F, SSB, and R within the 90% confidence intervals of the mean.

The Panel concluded that the WHAM extensions from the ASAP model formulation had little effects on stock status and scientific advice. The Panel considered Models 1-3 to involve minor modifications. Unlike other case studies, a WHAM model with NAA process errors was not presented for the Acadian redfish case study. Estimating these process errors prior to 1969 may be difficult for redfish because of the lack of age-composition information. This may affect the start year of a full WHAM model including cohort survival process errors on all ages.

A possible solution to the lack of age compositions is to include length composition information for years without age compositions using the growth-model branch of WHAM. Alternatively, this formulation could use length compositions for all years and condition age compositions. The Panel considers this to be useful research for the next management/research track. Related to this, state-space models may over-fit fishery age or length compositions in periods when there are no survey age/length composition information. This could be an issue for Acadian redfish if the start year of the assessment is considerably earlier than the start year of the spring and fall surveys. This should be investigated by varying the start year of the assessment.

The WG provided research recommendations that the Panel agreed with. These were:

1. The Acadian redfish assessment appears to have difficulty estimating NAA in the first year and historic (1913-1964) recruitment. Starting the model in a later year, when survey indices and age composition data are available, may improve estimation of these important parameters.
2. Explore the use of alternative distributional assumptions for the catch-at-age and survey index age compositions (e.g., logistic normal, Dirichlet). Using a different distribution may improve fit to the age composition, reducing the positive bias in the one step ahead residuals. In addition, doing so could improve fit to the survey indices, as was suggested by the re-weighting exercise in the 2023 management track assessment (NEFSC in prep.).
3. Explore a full state-space model configuration of WHAM (i.e., treating all NAA as random effects) for Acadian redfish. Treating all NAA as random effects may improve the fit to the survey indices by accounting for processes other than F that may be affecting stock abundance (e.g., migration between US and Canadian waters).

Georges Bank winter flounder case study

A virtual population analysis (VPA) has been used to assess Georges Bank (GB) winter flounder. The stock was last assessed in the 2022 Management Track with a rebuilding plan for a target

date of 2029. Many issues have been associated with the VPA-based assessment including major retrospective patterns and poor cohort tracking. A state-space framework has been suggested in the previous two assessments for this stock. The review panel from the last assessment recommended different recruitment assumptions for projections. In this case study, the WG explored fitting the 2022 assessment inputs used for the last VPA to WHAM. The WG used a stepwise approach to evaluate different WHAM configurations including 8 distributional functions for age compositions, 3 recruitment functions, 3 time-varying fisheries selectivities, and random effects on NAA. The WHAM configuration with the logistic normal distribution on all age composition and a 2dar1 process on numbers at age was identified as the most suitable model configuration using various criteria developed to evaluate the performance of a state-space model. Although a limited number of configurations was explored, this case study shows that moving to WHAM can improve the assessment. The proposed run has improved diagnostics compared to previous VPA runs and similar reference points and projections. The Panel supported a WHAM run for use in the 2025 Management Track assessment. However, the Panel suggests that more explorations need to be done before finalizing the model configurations for the management track. The Panel suggests that the WG consider the following recommendations:

- Adding random effects to life history and fishing processes one by one in a stepwise fashion may miss possible interactions among random effects assumed for different processes, the WG may consider to start from a WHAM configuration with full implementation of process errors and then reduce one-by-one to evaluate the model performance to identify the final WHAM configuration.
- Winter flounder are sensitive to changes in their thermal habitats and their distributions and life history are likely to be influenced by the climate-induced changes in their ranges. It may be important to consider time varying natural mortality and growth, possible shift in their distributions, and movement phenology. Random effects on natural mortality and growth may need to be considered. Time varying catchability for the surveys should also be considered, which may address issues of poor fitting on Canadian survey indices.
- The simulation self tests showed positive biases for recruitment, SSB and fishing mortality. More studies are needed to understand why these are all positive (SSB and F are likely to have biases of different directions).

Northwest Atlantic mackerel

Northwest Atlantic mackerel was historically assessed using a VPA model, which was rejected in 2000 due to a lack of convergence, survey variability and a strong retrospective pattern in SSB. During the mid-2000s, an ASAP model was accepted that indicated low F and high SSB, although the results also showed a lack of older fish in both fishery and survey catches, as well as the presence of significant retrospective patterns in SSB, F and recruitment. This ASAP model was later deemed not suitable for use in management and there was no accepted assessment model from 2009 until the 2017 benchmark. The 2017 benchmark assessment incorporated a new rangewide egg index. This model has a poor fit to the trawl survey Albatross index, indicating process variability that is not captured in the assessment model. Retrospective patterns

increased in magnitude in the 2021 and 2023 assessments, and an increase in SSB projected during the previous assessments was not realized. During the 2023 assessment, the review panel recommended the development of a state-space model to better deal with process variability and changing ecosystems.

The stock is composed of two spawning contingents, one that spawns in the southern Gulf of St Lawrence and another that spawns in southern New England, but during the winter months the two continents mix on the Northeast US shelf. The assessment data include ages 1-10+, one fishery from 1968-2022 and three survey series: Albatross trawl survey 1974-2008, Bigelow trawl survey 2009-2022, and egg survey 1983-2022. Catches peaked around 1970, with smaller peaks around 1990 and 2005. Commercial catch-at-age shows the cohorts behind the catch peaks.

The Panel evaluated an analysis of the 2023 assessment data, fitting a WHAM model which can be further enhanced for the 2025 management track assessment. The model is designed to closely resemble the ASAP model settings, with constant M=0.2 for all ages, time-invariant selectivity and random effects on NAA. The model diagnostics used consider convergence, residuals, AIC, retrospective patterns, prediction skill, estimation performance, and plausibility.

Different WHAM modeling options were explored for the age compositions, recruitment, time-varying selectivity, and the numbers-at-age process. The proposed model options involve logistic-normal-ar1-miss0 for the age compositions and a 2dar1 process for the numbers-at-age.

The proposed model does not have the best AIC, but it has good retro patterns. Overall, the WHAM model fits are comparable to ASAP, with good fits to the age composition. Retrospectives are better in WHAM but the prediction skill for the egg survey is poor. The bad fit to Albatross survey index still persists and is a known issue in this assessment. Compared to ASAP, the current development WHAM model has lower SSB in most years, except higher SSB around 1970, and higher F in many periods.

The Panel supported the ongoing model development, as well as the research plan that the WG had drafted, focusing on:

- exploration of high level of process error for numbers at age,
- time-varying survey catchability and selectivity,
- time-varying natural mortality, recruitment, and
- projection uncertainty.

The Panel also recommended focusing research on examining possible reasons why the model cannot follow the stock size trends in the log-transformed observed Albatross trawl survey indices, and considering what modeling options are appropriate for this data component. For example, analyze how the model results are affected if the magnitude of the observation error in the survey index is estimated rather than fixed. Another model option that could be considered is modeling time-varying survey catchability.

Appendix 1 - Terms of Reference for Applied State Space Models Research Track Stock Assessment

1. Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.
2. Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.
3. Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.
4. Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.
5. Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

Appendix 2 – Initial agenda for Applied State Space Research Track Assessment Peer Review meeting, February 12-15, 2024.

Applying State Space Models Research Track Assessment Peer Review Meeting

February 12-15, 2024

Meeting link: https://meet.google.com/fhd-msfm-pzz

DRAFT AGENDA* (v. 1/10/24)

**All times are approximate, and may be changed at the discretion of the Peer Review Panel chair. The meeting is open to the public; however, during the Report Writing sessions we ask that the public refrain from engaging in discussion with the Peer Review Panel.*

Monday, February 12, 2024

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:15 a.m.	Welcome/Logistics Introductions/Agenda/ Conduct of Meeting	Michele Traver, Assessment Process Lead Kristan Blackhart, PopDy Branch Chief Yong Chen, Panel Chair	
9:15 a.m. - 10:00 a.m.	Introduction/Executive Summary	Tim Miller (WG chair)	Review current use of state-space models in management, WG findings and recommendations
10:00 a.m. - 11:00 a.m.	TOR #5: GOM haddock	Charles Perretti	WP 5.1: Simple transition from ASAP to WHAM
11:00 a.m. - 11:15 a.m.	Break		
11:15 a.m. - 12:15 p.m.	TOR #5: GB winter flounder	Alex Hansell	WP 5.2: Simple transition from ASAP to WHAM
12:15 p.m. - 1:15 p.m.	Lunch		
1:15 p.m. - 2:15 p.m.	TOR #5: Redfish	Brian Linton	WP 5.3: Simple transition from ASAP to WHAM
2:15 p.m. - 3:15 p.m.	TOR #5: Mackerel	Kiersten Curti, Alex Hansell	WP 5.4: Simple transition from ASAP to WHAM
3:15 p.m. - 3:30 p.m.	Break		
3:30 p.m. - 4:00 p.m.	Discussion/Summary	Review Panel	
4:00 p.m. - 4:15 p.m.	Public Comment	Public	
4:15 p.m.	Adjourn		

Tuesday, February 13, 2024

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:05 a.m.	Welcome/Logistics/ Agenda	Michele Traver, Assessment Process Lead Yong Chen, Panel Chair	
9:05 a.m. - 10:45 a.m.	TOR #1	Tim Miller (WG Chair)	Miller et al WP1
10:45 a.m. - 11:00 a.m.	Break		
11:00 a.m. - 12:00 p.m.	TOR #1	Cheng Li	Li et al WP
12:00 p.m. - 1:00 p.m.	Lunch		
1:00 p.m. - 1:30 p.m.	Discussion/Summary	Review Panel	
1:30 p.m. - 2:30 p.m.	TOR #2	Tim Miller (WG Chair)	Miller et al WP1
2:30 p.m. - 2:45 p.m.	Break		
2:45 p.m. - 3:45 p.m.	TOR #2	Greg Britten, Liz Brooks	Britten et al. WP
3:45 p.m. - 4:00 p.m.	Public Comment		
4:00 p.m. - 4:30 p.m.	Discussion/Review/Summary	Review Panel	
4:30 p.m.	Adjourn		

Wednesday, February 14, 2024

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:05 a.m.	Welcome/Logistics/ Agenda	Michele Traver, Assessment Process Lead Yong Chen, Panel Chair	
9:05 a.m. - 10:15 a.m.	TOR #3: Environmental effects on recruitment	Greg Britten, Liz Brooks	Miller et al. WP2
10:15 a.m. - 10:30 a.m.	Break		
10:30 a.m. - 12:00 p.m.	TOR #3: Intro/Environmental effects on natural mortality	Tim Miller (WG Chair)	Britten et al. WP
12:00 p.m. - 1:00 p.m.	Lunch		
1:00 p.m. - 2:30 p.m.	TOR #3: Environmental effects on survey catchability	Amanda Hart, Alex Hansell	Hart et al. WP
2:30 p.m. - 3:15 p.m.	TOR #3: Reference points in stochastic populations	Tim Miller (WG Chair)	Miller WP
3:15 p.m. - 3:30 p.m.	Break		
3:30 p.m. - 4:00 p.m.	Discussion/Summary	Review Panel	
4:00 p.m. - 4:30 p.m.	TOR #4	Tim Miller (WG Chair)	
4:30 p.m. - 4:45 p.m.	Public Comment	Public	
4:45 p.m. - 5:15 p.m.	Discussion/Review/Summary	Review Panel	
5:15 p.m.	Adjourn		

Thursday, February 15, 2024

Time	Topic	Presenter(s)	Notes
9 a.m. - 9:05 a.m.	Logistics	Michele Traver, Assessment Process Lead Yong Chen, Panel Chair	
9:05 a.m. - 10:00 a.m.	Overview of panel findings	Review Panel	
10:00 a.m. - 12:00 p.m.	Report writing		
12:00 p.m. - 1:00 p.m	Lunch		
1:00 p.m. - 4:00 p.m.	Report writing		
4:00 p.m.	Adjourn		

Appendix 3 - Performance Work Statement (PWS) - Center for Independent Experts (CIE) Program – Applied State Space Models Research Track Peer 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 .

Scope

The Research Track Peer Review meeting is a formal, multiple-day meeting of stock assessment experts who serve as a panel to peer-review tabled stock assessments and models. The research track peer review is the cornerstone of the Northeast Region Coordinating Council stock assessment process, which includes assessment development, and report preparation (which is done by Working Groups or Atlantic States Marine Fisheries Commission (ASMFC) technical committees), assessment peer review (by the peer review panel), public presentations, and document publication. The results of this peer review will be incorporated into future management track assessments, which serve as the basis for developing fishery management recommendations.

The purpose of this meeting will be to provide an external peer review of the applying state space model framework. The requirements for the peer review follow. This Performance Work Statement (PWS) also includes: Annex 1: TORs for the research track, which are the responsibility of the analysts; Annex 2: a draft meeting agenda; Annex 3: Individual Independent Review Report Requirements; and Annex 4: Peer Reviewer Summary Report Requirements.

Requirements

NMFS requires three reviewers under this contract (i.e. subject to CIE standards for reviewers) to participate in the panel review. The chair, who is in addition to the three reviewers, will be provided by either the New England or Mid-Atlantic Fishery Management Council's Science and Statistical Committee; although the chair will be participating in this review, the chair's participation (i.e. labor and travel) is not covered by this contract.

Each reviewer will write an individual review report in accordance with the PWS, OMB Guidelines, and the TORs below. Modifications to the PWS and TORs cannot be made during the peer review, and any PWS or TORs modifications prior to the peer review shall be approved by the Contracting Officer's Representative (COR) and the CIE contractor. All TORs must be addressed in each reviewer's report. The reviewers shall have working knowledge and recent experience in the use and application of index-based, age-based, and state-space stock assessment models, including familiarity with retrospective patterns, model diagnostics from various population models, and how catch advice is provided from stock assessment models. In addition, knowledge and experience with simulation analyses is helpful.

Tasks for Reviewers

- Review the background materials and reports prior to the review meeting
    - Two weeks before the peer review, the Assessment Process Lead will electronically disseminate all necessary background information and reports to the CIE reviewers for the peer review.
- Attend and participate in the panel review meeting

    - The meeting will consist of presentations by NMFS and other scientists, stock assessment authors and others to facilitate the review, to provide any additional information required by the reviewers, and to answer any questions from reviewers
- Conduct an independent peer review in accordance with the requirements specified in this PWS and TORs, in adherence with the required formatting and content guidelines.
- Reviewers are not required to reach a consensus. Individual reviewer perspectives should be provided in their individual reports, and any lack of consensus should be clearly described in the panel's summary report.
- Each reviewer shall assist the Peer Review Panel Chair with contributions to the Peer Review Panel's Summary Report.
- Deliver individual Independent Reviewer Reports to NMFS according to the specified milestone dates.
- This report should explain whether each research track Term of Reference was or was not completed successfully during the peer review meeting, using the criteria specified below in the "Tasks for Peer Review Panel."
- If any existing Biological Reference Points (BRP) or their proxies are considered inappropriate, the Independent Report should include recommendations and justification for suitable alternatives. If such alternatives cannot be identified, then the report should indicate that the existing BRPs are the best available at this time.
- During the meeting, additional questions that were not in the Terms of Reference but that are directly related to the assessments and research topics may be raised. Comments on these questions should be included in a separate section at the end of the Independent Report produced by each reviewer.
- The Independent Report can also be used to provide greater detail than the Peer Reviewer Summary Report on specific stock assessment Terms of Reference or on additional questions raised during the meeting.

Tasks for Review panel

- During the peer review meeting, the panel is to determine whether each research track Term of Reference (TOR) was or was not completed successfully. To make this determination, panelists should consider whether the work provides a scientifically credible basis for developing fishery management advice. Criteria to consider include: whether the data were adequate and used properly, the analyses and models were carried out correctly, and the conclusions are correct/reasonable. If alternative assessment models and model assumptions are presented, evaluate their strengths and weaknesses and then recommend which, if any, scientific approach should be adopted. Where possible, the Peer Review Panel chair shall identify or facilitate agreement among the reviewers for each research track TOR.
- If the panel rejects any of the current BRP or BRP proxies (for BMSY and FMSY and MSY), the panel should explain why those particular BRPs or proxies are not suitable, and the panel should recommend suitable alternatives. If such alternatives cannot be identified, then the panel should indicate that the existing BRPs or BRP proxies are the best available at this time.

- Each reviewer shall complete the tasks in accordance with the PWS and Schedule of Milestones and Deliverables below.

Tasks for Peer Review Panel chair and reviewers combined:

- Review the Report of Applying State Space Models Research Track Working Group.

- The Peer Review Panel Chair, with the assistance from the reviewers, will write the Peer Reviewer Summary Report. Each reviewer and the chair will discuss whether they hold similar views on each research track Term of Reference and whether their opinions can be summarized into a single conclusion for all or only for some of the Terms of Reference of the peer review meeting. For terms where a similar view can be reached, the Peer Reviewer Summary Report will contain a summary of such opinions.

The chair's objective during this Peer Reviewer Summary Report development process will be to identify or facilitate the finding of an agreement rather than forcing the panel to reach an agreement. Again, the CIE reviewers are not required to reach a consensus. The chair will take the lead in editing and completing this report. The chair may express their opinion on each research track Term of Reference, either as part of the group opinion, or as a separate minority opinion. The Peer Reviewer Summary Report will not be submitted, reviewed, or approved by the Contractor.

Place of Performance
The place of performance shall be at NMFS Northeast Fisheries Science Center (NEFSC) in Woods Hole, MA, and via google meet link.

Period of Performance
The period of performance shall be from the time of award through February, 2024. Each 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.

MILESTONE	**DELIVERABLE**
Within 2 weeks of award	Contractor selects and confirms reviewers
Approximately 2 weeks later	Contractor provides the pre-review documents to the reviewers
February 12-15, 2024	Panel review meeting
Approximately 2 weeks later	Contractor receives draft reports
Within 2 weeks of receiving draft reports	Contractor submits final reports to the Government

* The Peer Reviewer Summary Report will not be submitted to, reviewed, or approved by the Contractor.

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
(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.

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

NMFS Project Contact
Michele Traver, NEFSC Assessment Process Lead
Northeast Fisheries Science Center
166 Water Street, Woods Hole, MA 02543
Michele.Traver@noaa.gov

Annex 1. Applying State Space Models Research Track Terms of Reference

1. Develop guidelines for diagnosing and selecting preferred state-space model structures. Comment on when alternative random effects assumptions and observation models are appropriate.
2. Investigate the efficacy of estimating stock-recruit functions within state-space models and their utility in generating scientific advice.
3. Develop guidelines for including ecosystem and environmental effects in assessment models and how to treat them for generating biological reference points and scientific advice.
4. Through simulation studies, evaluate relative performance of traditional and state-space models with respect to management metrics such as average and variability in catch, and stock and fishing mortality status. Consider factors such as life history type, sources of model-misspecification (as causes of retrospective patterns), and environmental effects.
5. Demonstrate any possible effects on stock status and scientific advice with incremental changes from statistical catch-at-age to full state-space model for applicable Northeast US stocks.

Research Track TORs:

General Clarification of Terms that may be used in the Research Track Terms of Reference

Guidance to Peer Review Panels about "Number of Models to include in the Peer Reviewer Report":

In general, for any TOR in which one or more models are explored by the Working Group, give a detailed presentation of the "best" model, including inputs, outputs, diagnostics of model adequacy, and sensitivity analyses that evaluate robustness of model results to the assumptions. In less detail, describe other models that were evaluated by the Working Group and explain their strengths, weaknesses and results in relation to the "best" model. If selection of a "best" model is not possible, present alternative models in detail, and summarize the relative utility each model, including a comparison of results. It should be highlighted whether any models represent a minority opinion.

On "Acceptable Biological Catch" (DOC Nat. Stand. Guidelines. Fed. Reg., v. 74, no. 11, 1-16-2009):

Acceptable biological catch (ABC) is a level of a stock or stock complex's annual catch that accounts for the scientific uncertainty in the estimate of Overfishing Limit (OFL) and any other scientific uncertainty…" (p. 3208) [In other words, OFL ≥ ABC.]

ABC for overfished stocks. For overfished stocks and stock complexes, a rebuilding ABC must be set to reflect the annual catch that is consistent with the schedule of fishing mortality rates in the rebuilding plan. (p. 3209)

NMFS expects that in most cases ABC will be reduced from OFL to reduce the probability that overfishing might occur in a year. (p. 3180)

ABC refers to a level of ''catch'' that is ''acceptable'' given the ''biological'' characteristics of the stock or stock complex. As such, Optimal Yield (OY) does not equate with ABC. The specification of OY is required to consider a variety of factors, including social and economic factors, and the protection of marine ecosystems, which are not part of the ABC concept. (p. 3189)

On "Vulnerability" (DOC Natl. Stand. Guidelines. Fed. Reg., v. 74, no. 11, 1-16-2009):

"Vulnerability. A stock's vulnerability is a combination of its productivity, which depends upon its life history characteristics, and its susceptibility to the fishery. Productivity refers to the capacity of the stock to produce Maximum Sustainable Yield (MSY) and to recover if the population is depleted, and susceptibility is the potential for the stock to be impacted by the fishery, which includes direct captures, as well as indirect impacts to the fishery (e.g., loss of habitat quality)." (p. 3205)

Participation among members of a Research Track Working Group:

Anyone participating in peer review meetings that will be running or presenting results from an assessment model is expected to supply the source code, a compiled executable, an input file with the proposed configuration, and a detailed model description in advance of the model

meeting. Source code for NOAA Toolbox programs is available on request. These measures allow transparency and a fair evaluation of differences that emerge between models.

Annex 2. Draft Review Meeting Agenda
{Final Meeting agenda to be provided at time of award}

Black Sea Bass Track Assessment Peer Review Meeting

December 5 – 7, 2023

For Details, Please see the following link: https://www.fisheries.noaa.gov/event/black-sea-bass-2023-research-track-peer-review

Annex 3. Individual Independent Peer Reviewer Report Requirements

1. The independent Peer Reviewer report shall be prefaced with an Executive Summary providing a concise summary of whether they accept or reject the work that they reviewed, with an explanation of their decision (strengths, weaknesses of the analyses, etc.).

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. The independent report shall be an independent peer review, and shall not simply repeat the contents of the Peer Reviewer Summary Report.

    a. Reviewers should describe in their own words the review activities completed during the panel review meeting, including a concise summary of whether they accept or reject the work that they reviewed, and explain their decisions (strengths, weaknesses of the analyses, etc.), 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 Peer Reviewer Summary Report that they believe might require further clarification.
    d. The report may include recommendations on how to improve future assessments.

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 4. Peer Reviewer Summary Report Requirements

1. The main body of the report shall consist of an introduction prepared by the Research Track Peer Review Panel chair that will include the background and a review of activities and comments on the appropriateness of the process in reaching the goals of the peer review meeting. Following the introduction, for each assessment /research topic reviewed, the report should address whether or not each Term of Reference of the Research Track Working Group was completed successfully. For each Term of Reference, the Peer Reviewer Summary Report should state why that Term of Reference was or was not completed successfully. It should also include whether they accept or reject the work that they reviewed, with an explanation of their decision (strengths, weaknesses of the analyses, etc.)

   To make this determination, the peer review panel chair and reviewers should consider whether or not the work provides a scientifically credible basis for developing fishery management advice. If the reviewers and peer review panel chair do not reach an agreement on a Term of Reference, the report should explain why. It is permissible to express majority as well as minority opinions.

   The report may include recommendations on how to improve future assessments.

2. If any existing Biological Reference Points (BRPs) or BRP proxies are considered inappropriate, include recommendations and justification for alternatives. If such alternatives cannot be identified, then indicate that the existing BRPs or BRP proxies are the best available at this time.

3. The report shall also include the bibliography of all materials provided during the peer review meeting, and relevant papers cited in the Peer Reviewer Summary Report, along with a copy of the CIE Performance Work Statement.

   The report shall also include as a separate appendix the assessment Terms of Reference used for the peer review meeting, including any changes to the Terms of Reference or specific topics/issues directly related to the assessments and requiring Panel advice.

Appendix 4 - Materials provided or referenced during the Applied State Space Research Track Stock Assessment Peer Review meeting

Working papers and presentations were available on a NEFSC website (https://apps-nefsc.fisheries.noaa.gov/saw/sasi.php) by selecting the species and year of assessment.

Working Papers and Background Documentation:

WP_1_Britten_Brooks_Miller_recruitment_functions_FINAL_withAppendix
WP_2_Hart_Hansell_environmental_effects_on_catchability_FINAL
WP_3_Li_et_al_Final
WP_4_Miller_et_al_factors_affecting_reliability_FINAL
WP_5_Miller_et_al_environmental_effects_on_M_FINAL
WP_6_Miller_stochastic_BRP_simulation_study_FINAL

WP_7_Monnahan_environmental_effects_on_growth_FINAL
WP_8_Hansell_GB_winter_flounder_final
WP_9_TOR5_Linton_Acadian_redfish_case_study_FINAL
WP_10_TOR5_Curti_Mackerel_FINAL
WP_11_TOR5_Perretti_GOM_haddock_case_study_FINAL
WP_12_TOR1_Liljestrand_et_al_degrees_of_process_error_FINAL

Presentations:

Intro_Day_1_Miller
TOR_1_Cheng Model Diagnostics
TOR_1_Miller
TOR_2_Miller
TORs_2_and_3_Britten_Brooks_Miller
TOR_3_Miller
TOR_3_Miller_WP_6
TOR_3_WP2_Hart_and_Hansell
TOR_4_and_draft_guidelines_Miller
TOR_5_GB_winter_flounder_Hansell_WP8_slides
TOR_5_Linton_Acadian_Redfish_FINAL
TOR_5_WP 10_Curti_Hansell
TOR_5_WP11_GOMHADDOCK_PERRETTI

Appendix 5 - Meeting attendees at the Applied State Space Models Research Track Stock Assessment Peer Review meeting

Applied State Space Models Research Track Peer Review Attendance
February 12-15, 2024

DFO - Fisheries and Oceans Canada
GARFO - Greater Atlantic Regional Fisheries Office
MADMF - Massachusetts Division of Marine Fisheries
MAFMC - Mid Atlantic Fisheries Management Council
NEFSC - Northeast Fisheries Science Center
SMAST - University of Massachusetts School of Marine Science and Technology
~~~~~~
Yong Chen - Chair
Anders Nielson - CIE Panel
Noel Cadigan - CIE Panel
Arni Magnusson - CIE Panel

Kristan Blackhart - NEFSC, Population Dynamics Branch Chief
Michele Traver - NEFSC, Assessment Process Lead
~~~~~~

Alex Dunn - NEFSC
Alex Hansell - NEFSC
Alicia Miller - NEFSC
Amanda Hart - NEFSC
Andrew Applegate - NEMFC staff
Angela Forristall - NEFMC
Anthony Wood - NEFSC
Brandon Muffley - MAFMC staff
Brian Linton - NEFSC
Burton Shank - NEFSC
Caroline Lehoux - DFO
Cate O'Keefe - NEFMC Executive Director
Catherine Foley - NEFSC
Chengxue li - NEFSC
Charles Adams - NEFSC
Charles Perretti - NEFSC
Chris Legault - NEFSC
Dan Hennen - NEFSC
Elisabeth Van Beveren - DFO
Emily Bodell - NEFMC staff
Emily Liljestrand - NEFSC
Gareth Lawson - Conservation Law Foundation
Gavin Fay - SMAST
Gregory Britten - Massachusetts Institute of Technology
Halle Berger - University of Connecticut Avery Point
Jackie ODell - Northeast Seafood Coalition
Jamie Cournane - NEMFC staff
Jessica Blaylock - NEFSC
John Wiedenmann - Rutgers University
Jon Deroba - NEFSC
Joseph Powers - Consultant
Jui-Han Chang - NEFSC
Kathy Sosebee - NEFSC
Kiersten Curti - NEFSC
Larry Alade - NEFSC
Libby Etrie - Conservation Law Foundation
Liz Brooks - NEFSC
Matt Cieri - Maine Department of Marine Resources
Melanie Griffin - MADMF
Mike Celestino - New Jersey Division of Fish and Wildlife
Peter Stephenson - Department of Fisheries Western Australia
Rachel Feeney - NEFMC staff
Robin Frede - NEFMC staff
Sam Truesdell - NEFSC
Sefatia Romeo Theken - Deputy Commissioner of Massachusetts Department of Fish and Game
Steve Cadrin - SMAST

Susan Wigley - NEFSC
Tara Dolan - MADMF
Tara Trinko Lake - NEFSC
Tim Miller - NEFSC
Toni Chute - NEFSC