Center for Independent Experts (CIE) independent peer review report for the SEDAR 65 HMS Atlantic Blacktip Shark Review

Anders Nielsen

January 2021

Executive Summary

The SEDAR 65 review of the assessment of Atlantic Blacktip Shark strengthened confidence in the assessment procedure used to assess the historic level of the stock, determine stock status, and provide projections. It is this reviewer's evaluation that the science reviewed constitutes the best scientific information currently available to assess the Atlantic Blacktip Shark stock.

Compared to the last assessment review (SEDAR 11, 2006) the data sources used and the assessment procedure have been updated. The catch of Atlantic Blacktip Shark is now divided into four components (longline, gillnets, others, and recreational). The assessment uses ten indices of abundance. The indices of abundance are selected/validated via an ICCAT inspired procedure, which includes/excludes data series based on a set of objective criteria. The procedure seems like a big step towards objectivity. The length-composition data are used where available and sufficient to inform the selectivity; for fleet-period combinations where it is not, the selectivity is borrowed from a similar fleet. Estimates of discard are not used in the base model. Important biological parameters (e.g., the steepness parameter in the assumed Beverton-Holt stock recruitment) are not estimated, but values are derived from life-history traits. Overall, the data decisions appear sound and robust.

The assessment framework has been updated to Stock Synthesis, which is a standard and well tested modelling framework. The main benefit for the current assessment is that it allows the use of all the different data sources in their natural form. The catches given in weight are included as weights, the recreational catches given in numbers are included as numbers, and the length compositions are included as is. This framework appears well chosen for this assessment.

The assessment does appear to be configured correctly with respect to (w.r.t.) the data sources and the choices w.r.t. biological parameters appear to be well reasoned. Hence the model is evaluated to be reliable to estimate historic stock levels, evaluate stock status, and provide stock projections. The Atlantic Blacktip Shark is concluded to not be overfished and not to be undergoing overfishing.

The details of the configuration include parameters that are fixed, prior distributions on other parameters, assumed variances or effective sample sizes, and indices which are smoothed across years. Such things – while not uncommon in assessment models –obstruct the model's ability to correctly quantify the uncertainties, so this reviewer does not consider the uncertainties presented (derived from the curvature of the objective function) as realistic. It would be recommended to use a different procedure to estimate the uncertainties (e.g., a bootstrap approach).

This and future assessments could benefit, and would be simpler to evaluate, if a standard set of model diagnostics were developed and provided. These could include: residuals (already provided, but should preferably be de-correlated), retrospective analysis, leave-out analysis, jitter-analysis, and simulation validation. Of special concern in this assessment was the problems w.r.t. the jitter-analysis. The jitter was not possible to complete for the base model, but was supplied for a restricted version, which reportedly gave indistinguishable results. For the restricted version it converged to local non-optimal solutions in 1/3 of the cases. It is therefore important to complete the jitter analysis for any subsequent assessment runs (future years) of this assessment to ensure that the estimates given are in fact based on a run that converged to the global minimum of the corresponding objective function.

Overall, the model and data have been substantially improved compared to previous assessments of Atlantic Blacktip Shark.

The review meeting was efficient and well organized by the SEDAR. However, having an assessment review online is not a good substitute for an actual review meeting. The discussion is slower, and hence fewer issues are raised. It is also not possible stand up and make an illustrative drawing where needed. Furthermore, the sharing of knowledge, which for other review meetings has been substantial (e.g., sharing tips and tricks of modelling, or introduction to new tools or software) does not happen if all breaks are in isolation. Having informal

discussions in person is much better for networking between assessment panel and reviewers, and overall makes the physical meetings more productive.

Background

The review workshop of the Atlantic blacktip shark assessment is part of the Southeast Data, Assessment, and Review (SEDAR 65) cooperative process for assessments conducted in NOAA Fisheries' Southeast Region. The meeting was conducted via five webinars (29 and 30 October, and 2, 4, and 5 November 2020) and a webinar pre-meeting some days prior (where the technical setup was tested and agenda briefly discussed). At the meeting, the assessment team (see appendix 3) presented the conclusions from the data workshop, all details of the assessment, stock status evaluation, and projections. In addition, the assessment panel answered all questions from the review panel (see appendix 3) and produced additional runs and model diagnostics requested by the review panel. The relevant documents (see Appendix 1) were made available in ample time prior to the meeting. The meeting was carefully prepared and well organized by SEDAR coordinator Kathleen Howington, who made the proceedings run very efficiently. The goal of such a review meeting is to strengthen confidence that the assessment is scientifically sound and that the results are reliable. The review panel, chaired by Dr. Beth Babcock, produced a joint consensus report. This report documents the independent review of CIE reviewer Anders Nielsen (see appendix 2 for the statement of work).

Description of this reviewer's role

This reviewer has independently read the assessment report, its appendices, and all supplementary documents deemed necessary in preparation for this review, participated in an online pre-meeting, participated actively in the online review meetings (29 and 30 October, and 2, 4, and 5 November 2020), identified key issues in the assessment and validation, suggested guidance, helped write the review panel's joint summary report, and independently authored this review report.

Findings regarding each term of reference

To ensure that all terms of reference are covered and that comments are interpreted with reference to the correct terms, the terms are listed (boldface) with corresponding reviewer comments following (standard font).

1. **Evaluate the data used in the assessment, including discussion of the strengths and weaknesses of data sources and decisions, and consider the following:**

    a. **Are data decisions made by the DW and AP sound and robust?**

The data decisions made by the data workshop (DW) and the assessment panel (AP) are sound and robust.

The fleets catching Atlantic Blacktip shark are divided into four components (longline, gillnets, others, and recreational). Recreational is the biggest component and consists of catches seen dead by an interviewer (A), catches reported dead (B1), and the part of the reported released catches (B2), which are estimated to have died due to the handling. In the text in the report the total recreational catches are denoted A+B1+B2-dead, which is a bit misleading, because the intended dash ("-") is easily mistaken for a minus symbol. A suggestion is to replace "B2-dead" with $B2_{dead}$ (using a subscript), or simply with 0.185*B2, because it is assumed that 18.5% of B2 dies. Estimated discards are not used in base model but included in a sensitivity run.

The indices of abundance are selected/validated via an ICCAT inspired procedure which considered many different criteria (e.g., spatial coverage, standardization, and index uncertainty quantification method). The procedure was very systematic, documented in working papers, and summarized in an understandable decision flowchart. Using this procedure made the selection process more objective.

A topic raised, in the assessment report and during the review meeting, was the conversion factor of 1.39 used to convert from dressed weight to whole weight. A plausible alternative conversion of 2.0, which is used by other agencies and had some support in a public comment, was tested as part of a sensitivity run.

The observed length compositions are sparsely available, so some fleets are set to borrow the length-based selectivity estimated from other fleets (without uncertainty). Furthermore, the fitted/predicted length compositions do not match the observations well (figures 3.3) within each year, but averaged over all years the predictions match the observations much closer (figure 3.4). The issues w.r.t. length compositions are common in fish stock assessment models, and it is an open research area how to best treat this part of the observations, so what is done here is clearly within standard practice, and as such is a well-tested approach.

    b. Are data uncertainties acknowledged, reported, and within normal or expected levels?

The data uncertainties are generally acknowledged, reported, and within normal and expected levels. However, the way uncertainties are propagated to the final quantitative estimates of interest is less straightforward. Many parameters are assumed fixed without uncertainties, recruitment deviates are penalized by a subjectively assigned variance parameter, catches are included in the model as exact, effective sample sizes are assigned, and part of the modelling process (smoothing of indices across years) is conducted external to the model. All of these things can prohibit the model from correctly following the uncertainties from observation to model results. In an ideal case, all important model parameters are estimated from the data, and the model is validated –

including the assumptions about distribution of the observations. Then the uncertainty in the observations is correctly propagated (via the estimation procedure) to the results of interest. This ideal situation is not seen often in assessment models, so in this regard this assessment is no exception.

The review panel raised questions about meaning of "inverse CV weighting" as explained in the assessment report (top page 62), where it could be interpreted as if the weighting was equal to the standard deviation (back-calculated from the CV), which would have been a mistake. This was cleared up by the assessment panel and the weighting is in fact inversely proportional to the variance, which is a standard statistical practice. It was suggested to adjust the text.

c. Are data applied properly within the assessment model?

The data are applied properly within the assessment model. The assessment model (Stock Synthesis) is designed with a great flexibility to accommodate different data types. Hence the data can be entered into the model pretty much "as is". This is one of the great strengths of the selected assessment model framework.

d. Are input data series reliable and sufficient to support the assessment approach and findings?

The input data series appear to be reliable and are sufficient to support the assessment approach and findings, but the jitter analysis (described under TOR 2) did reveal some difficulties. It is important to note that this assessment is not only based on the data series supplied as input to the model. As is often the case in assessments, the assessment panel needed to include information derived from other analyses and subjective choices (e.g., post release mortality rate, conversion factor between dressed weight and whole weight, recruitment deviation variance, and constant selectivity periods). In this reviewer's judgment the assessment panel made well-reasoned choices. Many of the choices were later explored by alternative choices in sensitivity runs. Based on overall evaluation of model predictions of observations (figures 3.2 and 3.4) it is concluded that the overall findings are supported. It strengthens this conclusion further that the included logistic sensitivity run showed the same overall results. Further confidence in the results could have been obtained by also applying a different model (simpler and requiring fewer ad-hoc choices, possibly only fitting to a subset of the data) and seeing similar overall results.

2. Evaluate and discuss the strengths and weaknesses of the method(s) used to assess the stock, taking into account the available data, and considering the following:

(This reviewer drafted the response to this TOR for the review panel's joint report. The response below is based on the original draft solely written by this reviewer and extended by this reviewer)

a. Are methods scientifically sound and robust?

Yes, the model is scientifically sound and robust. The model presented by the assessment panel for HMS Atlantic Blacktip Shark is the Stock Synthesis assessment model. Stock Synthesis is among the most applied stock assessment models in the US and in the world. It is part of the NOAA Fish and Fisheries Toolbox (Fish-Tools https://nmfs-fish-tools.github.io/). Stock Synthesis has been validated in numerous peer reviewed assessments (e.g., SEDAR 54: HMS Sandbar Shark, SEDAR 39: Atlantic Smooth Dogfish, and SEDAR 44: Atlantic Red Drum), in peer reviewed scientific journal papers (e.g., Method & Wetzel 2013, Punt & Maunder 2013, and Zhu et al. 2016), and in meetings dedicated to evaluate assessment models (e.g., World Conference on

Stock Assessment Methods for Sustainable Fisheries, 2013, Boston; Workshop on Recent Advances in Stock Assessment Models Worldwide, 2010, Nantes; and many Center for the Advancement of Population Assessment Methodology (CAPAM http://www.capamresearch.org/) workshops).

Stock Synthesis is one of the most general and complex assessment models, which is an advantage because it is applicable in many different scenarios and is able to accommodate many different types of observations. The many possible ways to setup and configure Stock Synthesis also increases the difficulty and knowledge required to operate the model correctly. It is therefore important to thoroughly validate the model implementation (configuration and data entry). The model for Atlantic Blacktip Shark was validated via standard (Pearson) residuals, which are not optimal for the multinomial distribution assumed for the length compositions and did show substantial patterns. It would have strengthened the confidence in the model implementation substantially if the main results and conclusions had been confirmed by comparing to an independent (simpler) model or if the main results had been compared to the previous model used for blacktip shark (ASPM). Such an analysis had been completed by the assessment team in a previous assessment of sandbar shark as a proof of concept and found that Stock Synthesis could be configured to be very similar to the ASPM.

An important model diagnostic is the so-called “jitter-analysis”. This is needed to validate that the model converged to its global minimum. Estimation in all such models are based on minimizing an objective function (negative log likelihood) in order to find the combination of model parameter values, which maximize the likelihood of the actual observations, within the constraints of the model constructed. In practice this is done by an iterative process. First initial values are supplied for all parameters, then a number of steps are taken where - in each step - the values are improved by following the gradient of the objective function to its globally lowest value. In models with a high number of nonlinear model parameters we need to verify that this worked – in many cases it will not. If the objective function has multiple local minimum values, then the minimization process could stop in one of those, if a boundary is encountered then the minimization could stop there, or if one (or more) parameter(s) can fully compensate for another parameter then it could stop at any combination. The jitter is a simple check that the minimization is resulting in the unique lowest value of the objective function, which is the correct estimate. The procedure works by starting in a number (often 100) of different initial values and verifying that in all cases the minimization process ends up in the same global minimum. A jitter analysis was requested by the assessment panel.

The assessment panel were unable to get the jitter analysis working for the configuration for the suggested base model, but they did produce a jitter analysis for a reduced version of the base model (Table 1), which the assessment panel reported produced indistinguishable results from the base model. The jitter analysis of the reduced model showed that in 1/3 of the runs the model converged to values which were not the global minimum. The good part was that in the remaining runs it did converge to the lowest value and that value was the same as for the original run of the reduced model, which reportedly produced indistinguishable results from the base model. So, confidence in the reported results has been strengthened. The unfortunate part is that when the model is updated with new data and re-run, then we will have an uncomfortably high probability (ca. 1/3 ?) that the model will not converge to the correct estimates. The only way to know if the convergence is to the correct estimates will be to a) rerun the jitter analysis for the corresponding reduced model and verify that the global minimum was obtained and b) verify that the reduced model still gives the same results as the base model.

Table 1: Adapted base model jitter results for global convergence

	Likelihood	Frequency
1	538.9[1]	67
2	539.4	1
3	540.3	3
4	540.4	1
5	540.5	1
6	540.6	1
7	540.8	1
8	540.8	1
9	541.3	1
10	541.5	2
11	541.7	2
12	541.8	1
13	541.9	1
14	542.0	1
15	542.4	1
16	543.0	1
17	545.5	2
18	545.6	2
19	546.7	1
20	548.0	1
21	576.3	1
22	707.2	1
23	846.2	1
24	1263.6	1
25	1297.5	1
	Total	97
[1] *Min*	*538.9*	

References:

Methot, Richard D. Jr., and Wetzel Chantell R. 2013, Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research.

Punt, André E., Maunder, Mark N. 2013. A review of integrated analysis in fisheries stock assessment. 2012. Fisheries Research.

Zhu, J., Maunder, M. N., Aires-da-Silva, A. M., Chen, Y. 2016. Estimation of growth within Stock Synthesis models: Management implications when using length-composition data. Fisheries Research

b. Are assessment models configured properly and consistent with standard practices?

The model has been configured properly and consistently with standard practices. In fact, the configuration options are in some cases inspired by already peer reviewed assessments (SEDAR 39: Smooth Dogfish and ICCAT Shortfin Mako assessment).

In broad strokes the configuration can be summarized by: a) Yearly catches in weight/numbers from four fleets are assumed known without error. b) Indices of abundance from ten fleets are assumed log-normally distributed with externally estimated CVs (Francis adjusted). c) Length compositions are assumed multinomially distributed with Francis or Harmonic mean adjusted effective sample sizes. d) Parametric selection curves are estimated if sufficiency length composition data are available, otherwise the selectivity is mirrored from an assumed similar fleet. e) The underlying population model is sex- and age-structured, with Beverton-Holt stock-recruitment (with penalized deviances), sex-specific Von Bertalanffy growth, and a common length-weight relationship.

The details of the configuration include parameters that are fixed, prior distributions on other parameters, assumed variances or effective sample sizes, and indices which are smoothed across years. Such things – while not uncommon in assessment models –obstruct the model's ability to correctly quantify the uncertainties.

c. Are the methods appropriate for the available data?

Yes, Stock Synthesis is capable of including data in its original format. The catches given in weight are included as weights, the recreational catches given in numbers are included as numbers, and the length compositions are included where available. One detail is that the length compositions are included as multinomial, which implicitly assume that compositions from a fleet within a year are negatively correlated, but the data most often show that such observations are positively correlated across neighboring length groups. This could affect the estimated uncertainties.

3. Evaluate the assessment findings and consider the following:

a. Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?

The estimates of abundance, exploitation, and biomass estimates are reliable, consistent with input data and population biological characteristics and useful to support status inference. This conclusion is reached, because the assessment is evaluated to be an acceptable description of the observations (see comments w.r.t. model validation under TOR 2). Spawning Stock Fecundity (SSF) is a relevant measure of spawning stock size, and so is the total fishing mortality (F=Z-M) for exploitation. The uncertainty estimates of these quantities (Table 3.10 and Figure 3.9) indicate that the model is able to estimate them with a similar accuracy for all years (this reviewer does not trust the absolute level of these uncertainty estimates, but relatively they can still be useful).

b. Is the stock overfished? What information helps you reach this conclusion?

The stock is not overfished, which is seen by the fact that the last year's Spawning Stock Fecundity (SSF) estimate is larger than its estimated reference point Minimum Stock Size Threshold (MSST) (Figures 3.9 and 3.10). The most recent ca. 10 years SSF estimates are however the lowest level of the entire time series, and the estimated confidence interval, which is likely too narrow, shows substantial probability mass (eyeballing ca 40% from Figure 3.9) of SSF being less than MSST, but a slight increase is seen in the last three years. The conclusion that the stock is not overfished is further supported by the logistic sensitivity analysis (Figure 3.b.11).

c. Is the stock undergoing overfishing? What information helps you reach this conclusion?

The stock is not undergoing overfishing, which is seen by the fact that the last year's estimate of total fishing mortality (F) is less than its estimated reference point F_{MSY} (Figures 3.9 and 3.10). This conclusion appears more certain then than the conclusion w.r.t. SSF>MSST, because in the last year the 95% confidence interval of F/F_{MSY} does not extend beyond one. Even if both confidence intervals are too narrow it appears that the conclusion w.r.t. undergoing overfishing is more certain than the conclusion w.r.t. overfished. The estimate of total fishing mortality in the most recent year is among the lowest observed in the entire time series. The conclusion that the stock is not undergoing overfishing is further supported by the logistic sensitivity analysis (figure 3.b.11).

d. Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?

The stock recruitment relationship is somewhat informative and useful for evaluation of productivity and future stock conditions. A Beverton-Holt stock recruitment curve is assumed in the assessment model with yearly (subjective) penalized deviations allowed. The steepness parameter is not estimated, but fixed at h=0.4, which is a value derived from life-history traits. Fixing the steepness parameter is not uncommon in assessment models and it does affect the later derived uncertainty estimates but deriving the steepness from life-history traits makes it more objective. A sensitivity analysis with fixed steepness at almost one (h=0.99) was presented and it did show a poorer fit (more systematic yearly deviations) (figure 1), which gives some support to the assigned value.

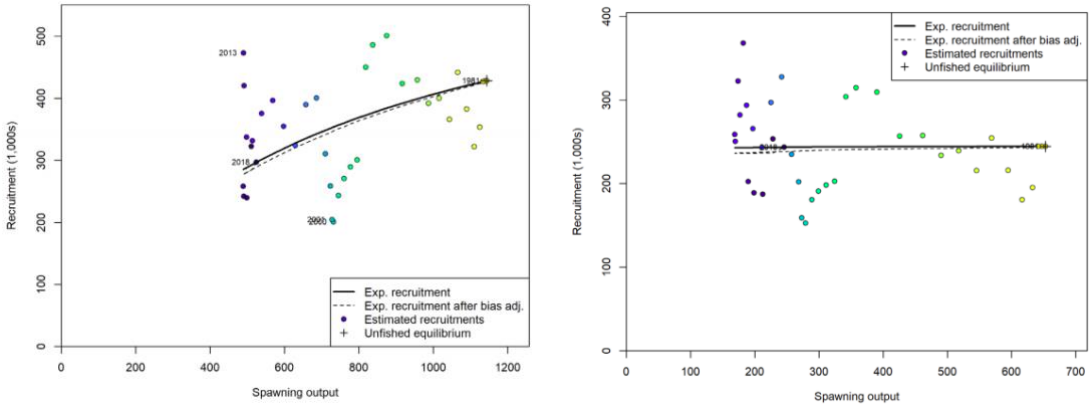

Figure 1: Stock recruitment with steepness h=0.4 (left fame) and with steepness fixed to h=.99 (right)

e. Are the quantitative estimates of the status determination criteria for this stock reliable? If not, are there other indicators that may be used to inform managers about stock trends and conditions?

The quantitative estimates of the status determination criteria for this stock are reliable and follows from the assumptions explained above.

4. Evaluate the stock projections, including discussing strengths and weaknesses, and consider the following:

a. Are the methods consistent with accepted practices and available data?

The stock projection method is within accepted practice and is consistent with the available data. The projection method is based on a projection method used for three similar peer reviewed assessments of blue shark (ICCAT), sandbar shark (SEDAR 54), and shortfin mako (ICCAT). The projection method is the standard catch-scenario projections which is a part of the Stock Synthesis assessment framework. However, to propagate uncertainties, the projection method applied uses maximum likelihood estimation and the delta method, which is different from the three previous assessments which used MCMC. Time constraints and less access to high performance computers (due to the Covid-19 crisis) are the reasons that the MCMC approach was not used. The delta method is a local linear normal approximation of whatever non-linearities are used in the forward propagation, so if the resulting distribution of the quantities of interest is far from a normal distribution, then the delta method will cause a bias. This bias was evaluated for mako and sandbar shark, where both methods were applied, and it was found that the delta method was slightly more pessimistic about the resulting catch-scenario, but that this bias was small compared to other uncertainties.

b. Are the methods appropriate for the assessment model and outputs?

The methods are appropriate for the assessment model and outputs. Catch-scenarios (0-200% in 10% steps) are projected forward for about two generation times, assuming selectivities within and among fleets similar to the last assessment year. The distribution of the future number of recruits is based on the estimated deviations, so that uncertainty is propagated forward, but it must be noted that those estimated deviations are penalized by a subjectively assigned deviance standard deviation.

c. Are the results informative and robust, and useful to support inferences of probable future conditions?

The results are informative, robust, and useful to support inferences of probable future conditions. The projections for different catch options are consistent with the assessment model used in the historic period. The key quantities used to determine stock status (spawning stock fecundity SSF and total fishing mortality F) were calculated for each projection and compared to the relevant reference points. This reviewer finds that this approach is a useful indication of the trends.

d. Are key uncertainties acknowledged, discussed, and reflected in the projection results?

Key uncertainties are acknowledged, discussed, and reflected in the projection results. The uncertainties projected forward are consistent with the uncertainties assumed in the assessment model. Model structure choices always affect how uncertainties in the observations are propagated to estimation/projection uncertainties. In this assessment model there are quite a few choices made, which directly influence the uncertainties (e.g., assigned fixed penalties on recruitment, fixed parameters in places where the estimation uncertainty became large, catches assumed known without error, external smoothing of recreational catch, assigned effective sample sizes). In addition, the multinominal does not adequately describe the correlation in composition data. So, in this reviewer's evaluation, the absolute levels of the projected uncertainties, are not reliable.

Some sensitivities (e.g., high catch and low productivity) were also presented to explore alternative states of the system, but to this reviewer these explorations are separate from the quantitative evaluation of uncertainties. Even if a high number of sensitivities were run, then they should not be considered as equally possible, and hence should not be evenly weighted. Assigning correct probability/weight to such runs would be difficult.

5. Consider how uncertainties in the assessment, and their potential consequences, are addressed.

a. Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods.

Uncertainties from the base model run are based on the standard in maximum likelihood estimation, which uses the curvature of the likelihood (inverse hessian) to describe the uncertainty of the estimated model parameters. Any derived quantities (functions of model parameters) are assigned uncertainty estimates via the delta-method (local linear approximation). This approach gives us (approximately) the estimated uncertainties of the model parameters and derived quantities. These uncertainty estimates are derived from the assumed or estimated error distributions of the observations. Uncertainties of the observations are simply propagated, via the model and estimation process, to uncertainties of the quantities of interest. This means that any mis-specification of the assumed distribution of the observations will lead to mis-estimation of the uncertainties. Notice, e.g., that if we assign a wrong variance parameter to a data source (e.g., a fleet) that does not necessarily mean that we will get wrong estimate values of SSF and F, but it will lead to wrong uncertainty estimates.

In the base model there are quite a few places where subjective choices will affect the estimated uncertainties. Fixed model parameter values, which are affecting the predicted values, are generally well reasoned, so it can be expected that the overall estimates of quantities relevant for management are reasonable. However, such fixed values obstruct the model's ability to correctly quantify the uncertainties:

- Prior distributions are assigned to model parameters. These are mostly "flat" and hence the influence of those will hopefully be minimal. However, this can be tricky, because a "flat" prior in one parametrization corresponds to a non-flat prior in a different parametrization. For instance, if a flat prior is assigned to a parameter 'log(a)', then that corresponds to assigning a non-flat prior distribution for 'a' itself. The term "non-informative" prior is often used, but really there is no such thing.
- Effective sample sizes are assigned and variance parameters are fixed. These assigned values directly affect the estimates of uncertainties returned from the model.
- Penalized deviances are used in the stock-recruitment relationship and the penalty term is assigned by assigning an externally fixed variance parameter. This acts as a zero-centered informative prior on the deviations.
- Several model parameters are fixed (not estimated from data). This means that whatever uncertainties would have occurred when estimating them are not included (or are absorbed elsewhere). In this context having fixed parameters are equivalent to having assigned highly informative priors. It is often the case in practical assessment models that some parameters cannot be estimated and are assigned fixed values (e.g., natural mortality, which is often confounded with fishing mortality). In such cases the consequences are explored by sensitivity analysis. It seems that the number of fixed parameters in this model is a bit higher than usually seen, and most of them are not explored by corresponding sensitivity runs. The values at which they are fixed appears well reasoned, so the estimates of the important stock and exploitation parameters can be correctly estimated, but the uncertainties cannot be expected to be accurately represented.
- Total catches are assumed to have negligible uncertainties, which is likely too optimistic, but sensitivity runs are conducted to illustrate the consequences.
- The multinomial distribution is used to describe composition data. The correlation structure in a multinomial distribution implies that, e.g., neighboring length-class observations are negatively correlated, but plots of observed and predicted compositions clearly shows positive correlations (figures 3.3).
- The external smoothing of the recreational catches effectively removes observation uncertainty before those observations enter the assessment model, which also will affect the assessment model's ability to provide uncertainty estimates.

In addition to providing the maximum likelihood derived estimates of uncertainties, some sensitivities (e.g., high catch and low productivity) were also presented to explore alternative states of the system. This does give some idea about different specific error scenarios, but there is no clear way of quantifying such results. It is not easy to judge if the sensitivity runs are representative, likely, or unlikely.

The assessment panel considered exploring the uncertainties via MCMC (but could not due to covid-19 related delays). MCMC is a good way to explore uncertainties if the main concern is non-normality of the noise distribution. However, MCMC is still only exploring the model as defined by the objective, so all the issues mentioned above would still apply.

An alternative to the MCMC would be bootstrapping. If we consider the model and estimation procedure simply as a means to obtain the quantitative estimates needed, then the error distribution of such estimates could possibly be approximated by a non-parametric bootstrap procedure, where observations (or residuals of observations) were re-sampled with replacement. This would produce a high number of pseudo data sets, and the estimates could be calculated for each.

b. Ensure that the implications of uncertainty in technical conclusions are clearly stated.

The procedure to characterize uncertainties and their implications in the technical conclusions is clearly stated.

6. Consider the research recommendations provided by the Data Workshop and Assessment Process and make any additional recommendations or prioritizations warranted.

(This reviewer drafted the response to this TOR for the review panel's joint report. The response below is based on the original draft solely written by this reviewer and extended by this reviewer)

a. Clearly denote research and monitoring that could improve the reliability of, and information provided by, future assessments.

The assessment team did explore the different ways of combining indices, as recommended from the data workshop. For age-0 the hierarchical Bayesian and dynamic factor analysis produced similar indices, so the latter was used. The inclusion in the assessment resulted in poor fit, non-convergence, or convergence to unreasonable parameter values. A subset of indices was used in a sensitivity analysis. This reviewer shares the assessment panel's conclusion that this could further be explored if more time was available.

This reviewer supports the assessment panel's own research recommendations, which include: a) Investigating ways to set up reproductive timing in Stock Synthesis (different versions) and to investigate sensitivities to different choices. This appears to be an important, but largely technical issue. b) Studying the effect of recreational management actions on the length compositions. c) Investigating different ways to parameterize selectivity. In addition to the suggestions by the assessment panel, which are simpler functions and more informed priors, a suggestion could be to look into formulations based on random effects (state-space models). This allows flexible models for selectivity with few model parameters by setting up processes (e.g., for F at a given length), then the only model parameters to be estimated are the level and standard deviation of the processes. A model based on this principle (Nielsen and Berg 2014) is routinely used in many ICES assessments and another such model has recently been developed at the Northeast Fisheries Science Center (https://github.com/timjmiller/wham). d) Investigating the proportionally few large sharks observed compared to the number of large sharks estimated to be in the population. This apparent dome-shaped selectivity can be caused by a number of different things including spatial distribution. It would be useful to report this "cryptic biomass" to monitor if it is (e.g.) increasing over time. Further, this also relates to flexible modelling of the selectivity (see c above). e) Improved model diagnostics. This and future assessments would benefit, and be simpler to evaluate, if a standard set of model diagnostics were developed and provided. These could include: residuals (already provided, but should be decorrelated), retrospective analysis, leave-out analysis, jitter-analysis, and simulation validation.

In addition, see a number of research recommendations under TOR 8 in this report and in the review panel's joint report.

References:

Nielsen, A. and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research 158, 96-101

b. Provide recommendations on possible ways to improve the SEDAR process.

The SEDAR process for this meeting was well organized. The meeting was efficient. The assessment panel was able to quickly answer questions and produce new runs and requested diagnostics. So, within the constraints imposed by Covid-19, this meeting was close to optimal. The support staff was excellent and very helpful.

The presentation team can help the review team by preparing focused presentations, as they are easier to follow (larger fonts, more figures, and less text) than on screen browsing of assessment reports.

Having an assessment review online is not a good substitute for an actual physical review meeting. The discussion is slower, and hence fewer issues are raised. Also, you cannot easily stand up and make an illustrative drawing where needed to explain an issue. Furthermore, the sharing of knowledge, which for other review meetings has been substantial (e.g., sharing tips and tricks of modelling, or introduction to new tools or software) does not happen if all breaks are in isolation. Having informal discussions in person is much better for networking between assessment panel and reviewers, and overall makes physical meetings more productive. Short term (for individual meetings) these compromises are necessary and tolerable, but long term, if the entire review process was shifted to be online, the quality of the reviews would suffer and the added benefits of the entire review process (sharing of practical scientific knowledge and networking) would be lost.

7. Consider whether the stock assessment constitutes the best scientific information available using the following criteria as appropriate: relevance, inclusiveness, objectivity, transparency, timeliness, verification, validation, and peer review of fishery management information.

The presented stock assessment constitutes the best scientific information currently available for Atlantic Blacktip Shark. The assessment panel has done a good job of including all relevant data sources and have set up objective criteria for inclusion of each data source (the ICCAT inspired check diagram). The model and assumptions are described well in the assessment report. The assessment model framework (Stock Synthesis) is thoroughly verified simply because it is so frequently used. Transparency could be improved, because the source code of the Stock Synthesis software is not (yet?) publicly available (to this reviewer's knowledge it is possible to request and get access to the source code if you have a valid scientific reason, but the code is not put forward for anyone to inspect). The validation of the model's ability to describe data, quantify uncertainties, and converge consistently could be improved (see TOR 2 and TOR 5, and suggestions for standard validations under TOR 8). From an international perspective, the peer review process appears very thorough.

8. Provide suggestions on key improvements in data or modeling approaches that should be considered when scheduling the next assessment.

In addition to the suggestions mentioned in the panel's joint report, this reviewer suggests that a more complete set of model validation / model diagnostics will be supplied. Such a set can include:

- Standardized residuals. This assessment already presented residuals, but for the multinomial distribution these are not de-correlated. This means that even if the observations were perfectly simulated according to the model, then the residuals would still not be independent. If we can't expect the residuals to be "good" when the observations correspond perfectly to the model, then it can be difficult to judge a poor fit from such residuals. It is in fact possible to construct de-correlated residuals from multinomial observations via the one-observation-ahead predictions (and by using that the multinomial density can be written as a product of successive binomials). The technique is described in Thygesen et al. (2017).
- Retrospective analysis. Since focus is often on the last year's estimate, it is important to verify that there is no last year systematic bias. An additional benefit of running the retrospective is that it often reveals if the model is unstable in some way. If, e.g., 2 out of 7 peals fail to converge, then the model is likely not sufficiently robust to recommend running for the next five years.
- Leave out runs. This analysis would strengthen confidence in robustness and furthermore reveal conflicting data sources and unduly influential data sources (e.g., fleets).
- Jitter analysis. A jitter analysis was presented for a slightly restricted version of the base model upon request of the reviewers (see under TOR 2). Is important to verify that a global optimum has been obtained and to verify stability w.r.t. initial values.
- Simulation analysis. Verifies that model parameters are identifiable, estimators are unbiased, and that confidence intervals have correct coverage.

It should be considered if the model can be simplified/restricted such that the jitter analysis always converges to the global minimum.

If the model is set up in a way such that the uncertainties derived from the objective function (inverse hessian or MCMC) cannot be expected to realistically describe the uncertainties of the derived estimates and derived quantities of interest, then consider exploring the uncertainties by bootstrapping.

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 (2), 317-339.

9. Prepare a Peer Review Summary summarizing the Panel's evaluation of the stock assessment and addressing each Term of Reference.

This report is the individual report of this reviewer. In addition, a summary report has been prepared jointly by the review panel. This reviewer contributed to all parts and specifically drafted the parts about the assessment model (TOR 2) and the parts considering the research recommendations (TOR 6).

Appendix 1: Bibliography of materials provided for review

Documents Prepared for SEDAR 65 Review Workshop			
SEDAR65-RW01	Updated Commercial Gillnet Length Composition Data for use in SEDAR 65	Dean Courtney, Alyssa Mathers, and Andrea Kroetz	9/18/2020
SEDAR65 RW02	Projections Conducted for the Atlantic Blacktip Shark Stock Synthesis Base Model Configuration at Alternative Fixed Total Allowable Catch (TAC) Limits	Dean Courtney	10/5/2020
Reference Documents			
SEDAR65-RD15	Marine Recreational Information Program Transition to Improved Survey Designs	John Foster and Kelly Denit	10/22/2020
SEDAR65-RD16	APAIS At-a-Glance	NOAA Fisheries, Marine Recreational Information Program	10/22/2020
SEDAR65-RD17	Field Procedures Manual: Access-Point Angler Intercept Survey	Atlantic Coastal Cooperative Statistics Program	10/22/2020
SEDAR65-RD18	National Marine Fisheries Service's Marine Recreational Information Program Survey Design and Statistical Methods for Estimation of Recreational Fisheries Catch and Effort	Katherine J. Papacostas and John Foster	10/22/2020
SEDAR65-RD19	Review of the Marine Recreational Information Program. .	The National Academies of Sciences, Engineering, and Medicine	10/22/2020
SEDAR65-RD20	Age-specific natural mortality rates in stock assessments: size-based vs. density-dependent	Joseph E. Powers	10/30/2020
SEDAR65-RD21	Modelling the effects of density-dependent mortality in juvenile red snapper caught as bycatch in Gulf of Mexico shrimp fisheries: Implications for management	Robyn E. Forrest, Murdoch K McAllister, Steven J.D. Martell, Carl J. Walters	10/30/2020

Appendix 2: A copy of this Performance Work Statement

Performance Work Statement (PWS)
National Oceanic and Atmospheric Administration (NOAA)
National Marine Fisheries Service (NMFS)
Center for Independent Experts (CIE) Program
External Independent Peer Review

SEDAR 65 HMS Atlantic Blacktip Shark 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 65 will be a CIE assessment review conducted for HMS Atlantic Blacktip Shark. 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 65 are the Atlantic stock of blacktip sharks in U.S. federal waters from Maine through Florida. 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**. The

tentative agenda of the panel review meeting is attached in **Annex 3** and the technical specifications required for this review are listed in **Annex 4.**

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 reviewers to have an expertise in shark population dynamics and/or shark assessments.

Tasks for Reviewers

1) Two weeks before the peer review, the 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 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) Additionally, two weeks prior to the peer review, the CIE reviewers will participate in a test to confirm that they have the necessary technical specifications provided in Annex 4 prepared in advance of the panel review meeting.

3) 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.

4) 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.

5) Each reviewer should assist the Chair of the meeting with contributions to the summary report. The Chair is not provided by the CIE under this contract.

6) Deliver their reports to the Government according to the specified milestones dates.

Place of Performance
The place of performance shall be online via gotowebinar.

Period of Performance
The period of performance shall be from the time of award through January 2021. 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.

Schedule	Milestones and Deliverables
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
October 29, 30 and November 2, 4, 5 2020	Panel will attend and participate in review webinars lasting approximately four and a half hours each day held between the hours of 8 am -8 pm CT
Approximately 3 weeks 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
Since this is a remote panel review, travel is neither required nor authorized for this contract.

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 for the Peer Review

SEDAR 65 Atlantic Blacktip Shark Assessment
Review Workshop Terms of Reference

Review Workshop Terms of Reference

1. Evaluate the data used in the assessment, including discussion of the strengths and weaknesses of data sources and decisions, and consider the following:
   a. Are data decisions made by the DW and AP sound and robust?
   b. Are data uncertainties acknowledged, reported, and within normal or expected levels?
   c. Are data applied properly within the assessment model?
   d. Are input data series reliable and sufficient to support the assessment approach and findings?
2. Evaluate and discuss the strengths and weaknesses of the method(s) used to assess the stock, taking into account the available data, and considering the following:
   a. Are methods scientifically sound and robust?
   b. Are assessment models configured properly and consistent with standard practices?
   c. Are the methods appropriate for the available data?
3. Evaluate the assessment findings and consider the following:
   a. Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?
   b. Is the stock overfished? What information helps you reach this conclusion?
   c. Is the stock undergoing overfishing? What information helps you reach this conclusion?
   d. Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?
   e. Are the quantitative estimates of the status determination criteria for this stock reliable? If not, are there other indicators that may be used to inform managers about stock trends and conditions?
4. Evaluate the stock projections, including discussing strengths and weaknesses, and consider the following:
   a. Are the methods consistent with accepted practices and available data?
   b. Are the methods appropriate for the assessment model and outputs?
   c. Are the results informative and robust, and useful to support inferences of probable future conditions?
   d. Are key uncertainties acknowledged, discussed, and reflected in the projection results?
5. Consider how uncertainties in the assessment, and their potential consequences, are addressed.
   a. Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods.
   b. Ensure that the implications of uncertainty in technical conclusions are clearly stated.
6. Consider the research recommendations provided by the Data Workshop and Assessment Process and make any additional recommendations or prioritizations warranted.

a. Clearly denote research and monitoring that could improve the reliability of, and information provided by, future assessments.
b. Provide recommendations on possible ways to improve the SEDAR process.

7. Consider whether the stock assessment constitutes the best scientific information available using the following criteria as appropriate: relevance, inclusiveness, objectivity, transparency, timeliness, verification, validation, and peer review of fishery management information.
8. Provide suggestions on key improvements in data or modeling approaches that should be considered when scheduling the next assessment.
9. Prepare a Peer Review Summary summarizing the Panel's evaluation of the stock assessment and addressing each Term of Reference.

Annex 3: Tentative Agenda - SEDAR 65 Atlantic Blacktip Shark Assessment Review
Via webinar

October 29 - November 5, 2020

Each day will consist of a 4.5 hour long webinar held between the times of 8 am and 8 pm CT
The start and end times of each webinar are dependent on CIE and analyst availability

October 29- Introductions and Opening Remarks		Coordinator
	- Agenda Review, TOR, Task Assignments	
	Assessment Presentations	Dean Courtney
October 30 – Assessment Presentation continued		Dean Courtney

October 29 and 30 Goals: Initial presentations completed, sensitivities and modifications identified.

November 2 -	Panel Discussion	Chair
	- Review additional analyses, sensitivities	
	- Consensus recommendations and comments	Chair

November 2 Goals: Final sensitivities identified, preferred models selected, projection approaches approved, Summary report drafts begun

November 4 - Panel Discussion		Chair
	- Final sensitivities reviewed.	
	- Projections reviewed.	
November 5 Panel Discussion or Work Session		Chair
	- Review Consensus Reports	

November 4 and 5 Goals: Complete assessment work and discussions. Final results available. Draft Summary Report reviewed.

Annex 4: SEDAR 65 HMS Atlantic Blacktip Shark Review workshop minimum technical requirements

1. Computer
2. Microphone and speakers (headset recommended)
3. GoToWebinar desktop app (JavaScript enabled) available for download here: https://support.goto.com/webinar/help/download-now-g2w010002
4. Internet: 1 Mbps or better (wired preferred)
5. Web browser:
    a. Google Chrome v57 or later
    b. Mozilla Firefox v52 or later
    c. Internet Explorer v10 or later
    d. Microsoft Edge v12 or later
    e. Apple Safari v10 or later
6. Operating system
    a. Windows 7 - Windows 10
    b. Mac OS X 10.9 (Mavericks) - macOS 10.15 (Catalina)
7. 2GB of RAM (minimum), 4GB or more of RAM (recommended)
8. Smart phone for use as audio backup and internet hotspot (recommended)

Appendix 3: List of participants

Review Panelist

Beth Babcock	Chair	University of Miami: RSMAS
Anders Nielsen	CIE	DTU-Aqua Technical University of Denmark
John Neilson	CIE	Independent fisheries Scientist
Joe Powers	CIE	Joseph Powers Consulting

Analytical Representatives

Dean Courtney	Lead Assessment Representative	NMFS: HMS
Xinsheng Zhang	Assessment Representative	NMFS: HMS
Enric Cortes	Assessment representative	NMFS:HMS

Council and Agency Staff

Kathleen Howington	Coordinator	SEDAR
Karyl Brewster-Geiz	HMS Management	NMFS: HMS
Clifford Hutt	HMS Staff	NMFS: HMS
Heather Baertlein	HMS Staff	NMFS: HMS

Review Workshop Attendees

Catherine Puma	Observer	University of Miami
Chip Collier	Observer	SAFMC
John Carlson	Observer	NMFS
Julie Neer	Observer	SEDAR
Manoj Shivani	Observer	NTVI Federal
Rusty Hudson	Observer	DSF