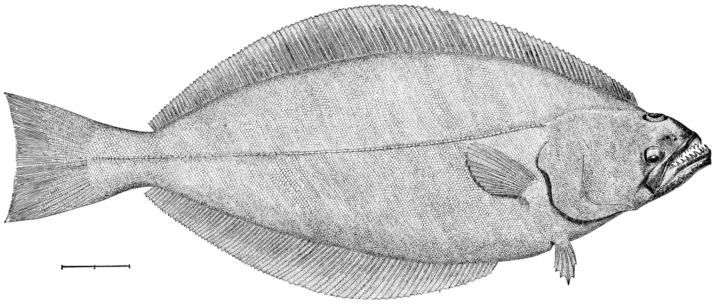
Center for Independent Experts (CIE) Independent Peer Review Report of the **Gulf of Alaska Northern and Southern Rock Sole, Bering Sea Aleutian Islands Greenland Turbot Assessments,**

Seattle and online, April 5-9th 2021

Sven Kupschus (Sven@kupschus.net)



THE GREENLAND TURBOT.

Prepared for:

Center for Independent Experts (CIE) Independent System for Peer Review

Executive Summary

The panel was asked to review the latest stock assessments conducted for Bering Sea and Aleutian Islands turbot and northern and southern rock sole in the Gulf of Alaska. Both rock sole stocks use the same data sources and methodologies as well as showing a large degree of synchrony in their stock dynamics so they were essentially treated as a single discussion.

I concluded that despite some not as yet resolved issues in all assessments they are suitable for management advice as provided under the tier 3a classification and on the evidence presented are the most scientifically appropriate models to apply.

Option for future improvements were discussed based on the diagnosis of the current models and some preliminary analyses were presented which would require further quality control before they could be considered improvements to the models used in management.

The assessments all suffer from some issues in modelling growth but it appears the reasons for this varies by stock and the approaches used to model it within the assessment. The issues estimating size-at-age (mean and/or variance) ultimately make it more difficult to estimate selection parameters appropriately. However, for turbot I consider the greater need to deal with the overly complex selection patterns and time blocks as a greater priority while for the rock sole assessments I would judge the growth to play a bigger part in the poor fit to the length distributions.

Biomass estimates are within safe biological limits for all stocks and the stocks are not overfished. Exploitation levels are higher in the turbot stock than in either of the rock sole stocks and the quota is considered restrictive so from a management perspective I would judge the turbot assessment to be a higher priority for improvements than the rock sole stocks. In addition, it seems easier to make more rapid progress on this given the data and modelling approach.

Table of Contents:

Executive Summary....2

Background....5

Summary of findings Bering Sea and Aleutian Islands Greenland turbot....6

Prioritization of research....6

Review activities Bering Sea and Aleutian Islands Greenland turbot....8

Growth:....8

Catch:....9

Longline survey index:....11

Catchability and selectivity:....12

Time blocks on selectivity:....12

Selectivity functions:....14

Sex off-sets for selectivities:....15

Survey selectivities and catchabilities:....16

Other discussions:....18

Summary of findings Gulf of Alaska Northern and Southern Rock Sole assessment....20

Prioritization of research:....21

Review activities Gulf of Alaska Northern and Southern Rock Sole assessment....21

Catch data:....21

Growth:....22

Surveys:....24

Selectivity:....27

Lack of contrast:....28

Model consistency:....29

Critique of Review Process....30

Appendix 1: Bibliography of materials provided for review....31

CIE Materials:....31

Appendix 2: A copy of this Performance Work Statement....32

External Independent Peer Review....32

Gulf of Alaska Northern and Southern Rock Sole,....32

Bering Sea Aleutian Islands Greenland Turbot....32

Bering Sea and Aleutian Islands Greenland Turbot....33

Annex 1: Peer Review Report Requirements....36

Bering Sea and Aleutian Islands Greenland turbot....37

Gulf of Alaska rock soles 37
Annex 3: Tentative Agenda..38
CIE Panel Review of Gulf of Alaska Northern and Southern Rock Sole, ...38
Bering Sea Aleutian Islands Greenland Turbot...38
Appendix 3: Panel membership or other pertinent information from the panel review meeting. 39
References: 39

Background

The review of the Bering Sea and Aleutian Island Greenland turbot and Gulf of Alaska rock soles was conducted by video link from the 5th-9th of April 2021. The meeting was chaired by Kalei Shotwell and choreographed by the assessor Meahgan Brian, also the assessor for the two stocks under review. The independent reviewers provided by the CIE were Anders Nielsen, Collin Millar, and myself.

The first two days were scheduled for review of the turbot assessment starting with a discussion with data collectors (Lyle Britt, John Brogan, Raul Ramirez, Kevin Siwicke) and transitioning to a presentation by the assessor of the assessment to be reviewed. The remainder of the time was spent examining details of the assessment, its strengths and weaknesses and discussing alternate ways to interpret the data. The process was repeated on the subsequent two days for the rock sole assessment with the relevant data collectors (Daniel Armellino, John Brogan, Wayne Palsson).

The final day was available to review output in response to some initial reviewer requests, discussing follow up topics from the reviewers and summarising / prioritising the reviewers' recommendations, but not all the time allocated to these tasks was necessary to conclude the review. The meeting chair kindly provided the document list and attendance list (Appendix 1,3) for the purposes of this report.

In preparation for the meeting, I read the documentation made available two weeks in advance of the meeting and listened to the recorded presentations on the data collection and scientific research conducted. I found the presentations highly informative and together with the discussion opportunity with presenters at the meeting they allowed me to develop a feel for the ecology, the fishery and the relevant topics to consider for the assessment review process. As someone who vies the assessments as a balancing of the different data under the consideration of different mechanisms and processes this is an essential part of the review process for me. The description of the sampling designs and data summaries allowed me to assess the information content and the critical stock trends suggested by the data in conjunction with the assessment presentation. I would at this point specifically like to commend the presenters, their openness and efforts to respond to questions. It also set a good tone for the remainder of the review. In fact, the preparation and the entire meeting itself were marked by open and free exchanges of ideas and information between the assessor and the reviewers despite what is designed to be a 'critical' review. Interventions by the chair were merely necessary to keep the discussion in time to ensure that all terms of reference were addressed and to manage the challenges of a virtual meeting.

The documents provided were mostly the standard assessment reports and while they did demonstrate the factual evidence for the current assessment, it was not always possible to follow the reasoning for certain decisions being made. Mostly this related to historic decisions, and the assessor with support from Jim Ianelli was able to address many of these concerns during the presentations and discussions as well as provide more information on the historic activities of the fishery and their responses to changes in management. I very much appreciated the time taken and the willingness to take on board different ideas and to take any criticisms as constructive ideas for improving the assessments. Some of these aspects were particularly challenging in this review because both assessments were conducted by the same assessor, so I consider the cooperative approach taken and the effort made as particularly noteworthy.

Summary of findings Bering Sea and Aleutian Islands Greenland turbot

The current statistical catch-at-age assessment methodology implemented in SS3 was developed in 2018 (Bryan *et al.*, 2018) with the 2020 assessment merely updating the data with the most recent catches and survey values. The stock is currently managed as part of the ‘deepwater flatfish complex’ under Tier 3a rules having commendably developed historically from Tier 4 biomass index-based methods.

The model provides a sound scientific basis for the provision of advice using the most advances scientific knowledge available. The available data are capable of describing the dynamics of the stock within the assessment model used. Improvements in the model parameterisation are always possible but will not substantially alter the management of this stock given the likely magnitude of their impacts. Furthermore, such refinements will always have risks as well as benefits, the relative value of which is likely to be influenced by individual perspective and experiences. The resources in terms of assessor time and data collection must also be considered at a time where these are limited. In this context, I conclude that the 2020 assessment is the best available evidence on which to base the advice.

The improvements in knowledge of the stock dynamics associated with the more complex models facilitate greater yield from the fishery while better ensuring the sustainability. The increasing parameter complexity in light of few recent improvements in the available information data sources does represent some issues with model stability as mentioned in Bryan (2018). In addition, many of the decisions on how the data is used in the assessment are historic and at least not regularly revisited to examine whether the original reasons for those choices are still relevant or appropriate in the context of the latest understanding of the biological dynamics or recent developments in methodology and assessment practices. With a naïve perspective and an independent view, the CIE panel was able to suggest some potential avenues for improvements to the model and the data collection based on a review of the 2020 assessment. However, there was insufficient time to scientifically evaluate the effects of such improvements, so it is not possible to conclude that these would present better approaches or provide better advice. The review attempted to prioritise the need for such investigations in the context of this stock and its management requirements, but will also require prioritisation across all the stock assessment responsibilities which is beyond the scope of this review.

Prioritization of research:

I prioritised the assessment needs as follows:

1) Highest priority for the Greenland turbot assessment is a re-evaluation of the highly parameterised selectivities in this assessment using multiple time blocks and complex selectivity functions which in some cases produce unexpected results from the perspective of the processes being modelled and can have deleterious effects on the ability to effectively manage the stock under some future stock status conditions. It is recommended to start from a much simpler selectivity model and add complexity stepwise critically evaluating the evidence base for added complexity in the context of the information available from the data sources and not necessarily based only on the assessment diagnostics. Alternatively, it is possible to examine the model stability in respect to the selectivity parameters from MCMC and jitter analysis and apply a reductionist approach to the complexity issue. However, this is likely to be more time consuming and more difficult to prioritise due to the interactions between the parameters within the model.

2) Although a catch-at-age model the assessment derives its information on relative cohort strength predominantly from length data using the age data only to determine mean size-at-age. These are modelled using a single constant growth function for the entire timeseries. Appropriate specification of the growth is vital for the assessments ability to correctly assess recruitment and mortality dynamics. It seems in the assessment the calculation of growth is based on the available age data from the survey without considering the length stratified random sampling used in the age collections. A stratified means approach is likely to produce less biased estimates of size-at-age. One could argue this should be highest priority as any changes are likely to impact the evaluation of selectivities and this would be relatively simple to estimate. However, it is not necessarily easily implementable within the SS3 frame work and is likely to predominantly affect the abundant younger cohorts which are more distinctly identified based on their length irrespective of small biases. If this is to be done it should be done prior to examination of the selectivities.

3) While model stability is a critical aspect of an assessment from a management perspective, further investigations regarding model stability based on the 2020 model are highly likely to be subjective to the final choice of model and would have to be repeated were changes to the model to be derived from 1 and 2. There was substantial discussions on methods used to investigate model stability, but my view of the discussions was more to diagnose where the model is having difficulties and where opportunities existed for improvements rather than as an evaluation of the suitability of the existing model for management. This was mainly because there was insufficient time to fully diagnose and compare multiple MCMC and jitter analysis runs.

4) The catch data in the early timeseries seem to comparatively uncertain as noted already by the 2007 CIE review. Some of this relates to the certainty around the international data particularly the species splits which were not reported separately. However, near coincidental with the commencement in 1986 of the current observer program, there is also a sharp drop in the estimated catches used in the assessment. It is not clear. The model interprets this as a sharp decline in abundance at roughly constant F. To balance its population, it assumes two exceptionally large recruitment pulses (early 60's and mid 70's) at a time when there is no compositional data in the assessment. These cohorts have exited the fishery but they still play an important role in deriving management metrics through their impacts on current estimates of selectivity and F. The magnitude of this effect is unknown because of the complex time block pattern.

   Not only are estimated recruitment pulses of lower magnitude, but average recruitment is estimated to have been much lower since the inclusion of compositional data. These estimates greatly influence the perception of stock productivity and presumably management reference points.

   Answers to this question may be obtainable through the collected otoliths that apparently exist from the fishery at a time when these cohorts should still have been present in the fishery. While it may not be straight forward to include this information in the assessment, it does provide a critically important test of the ability of the existing model to reflect the stock dynamics.

5) The assumption of steepness = 0.79 (from meta-analysis) appears to be necessary for this assessment due to a lack of a discernible stock recruitment relationship. Aside from the inherent autocorrelation in SSB there is some autocorrelation in recruitment. There are a number of possible reasons for this:
   a. Environmental ecosystem effects.
   b. Ageing difficulties as the age and growth team suggested the species was difficult to age so that abundant cohorts may be attributed to adjacent cohort through an imprecise growth function.
   c. Density dependent variation in growth, again resulting in the reassignment of a strong cohort to adjacent cohorts.

Given the evidence for recent poor recruitments by the shelf survey it can be anticipated that this question will in the medium-term become an issue for management, so collecting the evidence and exploring options now will pre-empt poorly informed management decisions in future. This is particularly relevant given the reported uncertainty of a return of the slope survey that plays an important part in this assessment to monitor the abundance trends of these cohorts as they move out of the area covered by the shelf survey.

Review activities Bering Sea and Aleutian Islands Greenland turbot

Growth:

Growth is currently incorrectly estimated within the model because the mean weight at age from the longline survey does not reflect the length stratified sampling regime. The age information as used is representative of the age samples, but these represent a biased subset of the length samples with abundant length usually underrepresented in the age information due to a fixed maximum number of otolith samples per length group. The magnitude of this bias is strongly dependent on the difference in the proportion of samples taken from each length group (which I was unable to get a clear idea of from the survey presentation), and the relationship of the mean size-at age to the underrepresented length classes. A weighted mean should be used. Additionally, there is a need to include the currently ignored variance derived from the variation in length samples. This is likely to increase the uncertainty in mean size-at-age but also help the model deal with what seem to be overly influential length information (large standardised residuals) which seem to extend beyond the size the individuals grow to in the current specification. A trade-off may be that the model is less able to infer age from length (down weight the compositional information) but their good cohort signals is apparent in the length information and a re-specification may resolve model conflict by reassigning these individuals to the appropriate cohort.

Initial examination of the mean length-at-age data (Figure 1) to me suggested some step-change in growth with a sudden decrease in the size-at-age. This was also picked up by the model in its residuals with a more general cohort blocking of positive and negative residuals (Figure 1). This could potentially be an artefact of the estimation of mean growth as just prior to the blocks of positive residuals some recruitment pulses moved through the population. Alternatively environmental or ecological factors may be responsible.

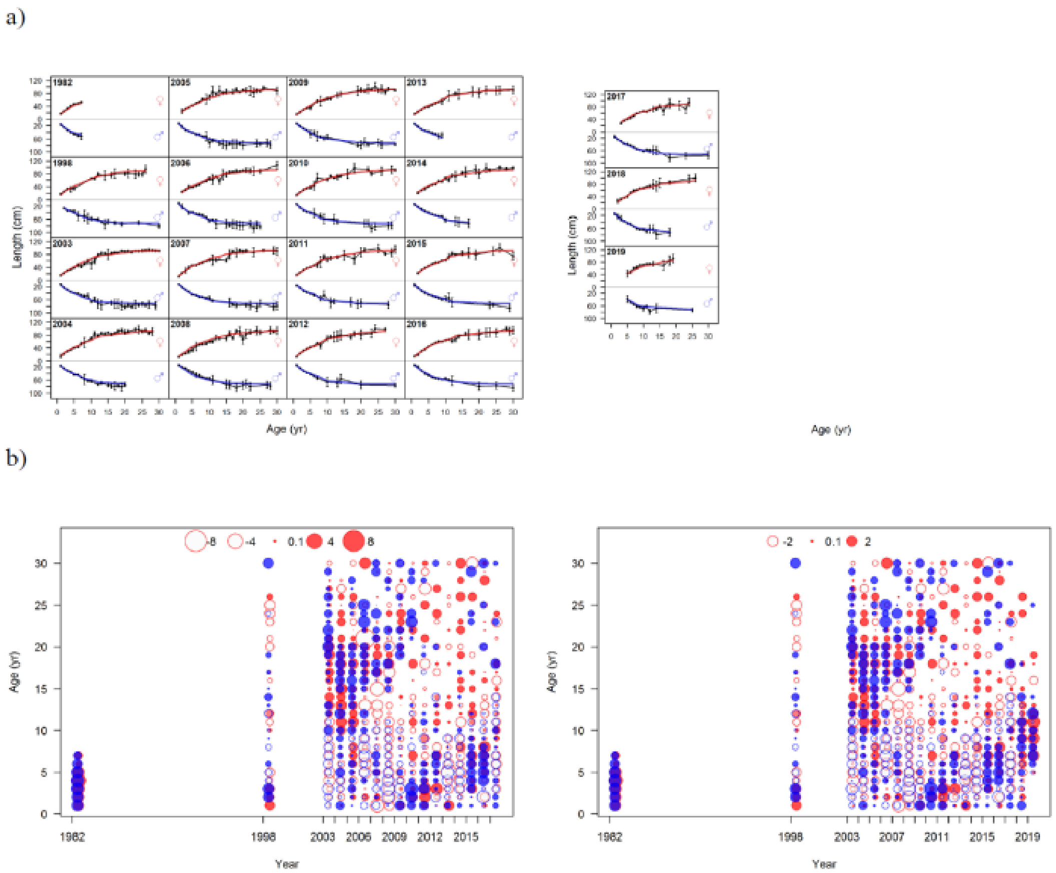


Figure 1 a) Length at age data and fit (females - red line, males – blue line) by Model 16.4a (2020) and b) the standardized residuals from Model 16.4 (left) and Model 16.4a (2020) (right). The closed bubbles are positive residuals (underestimation) and open bubbles are negative residuals (overestimation). Red bubbles are female and blue are male.

The magnitude in the change in size-size at age is small compared to the annual growth interval for young individuals and for these ages it is unlikely to preclude the appropriate interpretation of the length frequencies. Past maturity there is relatively little information on age from the lengths so the model needs to have picked up the recruitment signal previously to interpret the adult length information correctly. Therefore, I conclude that the assumption of a constant growth pattern only minorly impacts the model's ability to appropriately interpret the length frequencies, but this should be revisited if and when a more appropriate method for the interpretation of the mean-size-at-age appropriately accounting for the age sampling design is implemented. Given the strength of the recruitment signals in the data and the assessment, the concern is of a lower priority for this stock.

Catch:

My review of the individual data sources suggests that these are suitable for the assessment of stock dynamic trends in the context of a catch-at-age assessment. Unfortunately, the historic data (1960-1985) are comparatively uncertain in terms of the magnitude of catches due to foreign fishing activity and depauperate in compositional and abundance index information. Provided that the relevant stock dynamic parameters can be estimated based on the recent period and have remained constant over the entire period this does not necessarily suggest an assessment approach for the entire timeseries is not possible. To me it does suggest that it would be important to conduct some sensitivity analyses to examine the potential impact of the data characteristics on future management advice. Such examinations should include options with different assessment start dates or differential treatment of the uncertainty in catches (preferably bias as well as than variance).

My reasons are:

- The current model interprets the greater catches prior to the mid 80's as having been the results of two extraordinary recruitment pulses and a significantly higher background recruitment level, with virtually no discernible change in F. In contrast one would have to assume *a priori* that the exclusion of the substantial foreign fleet would have to have led to a reduction in effort which should be reflected in F unless the US fleet was able to take up the effort instantly. This would have been despite the commensurate steep decline in catches which seems unlikely from an economic perspective. While such coincidence is certainly possible it is necessary to examine if there is direct evidence available. At the meeting it was suggested that a substantial number of unaged otoliths from the fishery exist. Some of the early samples should according to the current model output still contain individuals of those strong year-classes which could provide evidence of the relative recruitment strengths as well as an indication of relative F. While it may not be possible to formally include this data in the assessment if only the largest / most informative individuals were aged it would provide a means to evaluate the credibility of the different assessment model options.
- In the absence of direct evidence for the appropriateness of the current model interpretation a sensitivity analysis on the likely impact on management advice, particularly the ability to estimate steepness (currently fixed) and selectivities (some currently poorly defined, see later section) should be examined. Options could include reduction of weighting of the historic catches, later assessment start dates, catch multipliers on historic data, etc.
- To consider the potential impact of variation in growth I was provided the size-at-age information used to determine the mean size-at-age for turbot used in the assessment. I proceeded to model growth based on an interaction (surface spline) between cohort and age to detect changes in cohorts over time. The data did not contain the necessary information on the proportion of individual sampled-at-length so cannot attribute the appropriate weighting. It is therefore it is consistent with the current size-at-age calculation and directly comparable. The analysis presented but not shown here for brevity concluded that there was a significant interaction between age and cohort. Prediction from the model indicated though significant these effects were relatively small compared to the annual growth increment at younger ages. Data at older ages seemed to be generally quite variable but the comparatively rare samples made a coherent analysis difficult. I concluded that at the older ages length is a relatively poor indicator of age precluding the model from separating cohorts based on length data only and therefore unlikely to suffer from the assumption of a single growth parameterisation.
- Unfortunately, the question is closely tied to the question of catchability as indicated by a preliminary attempt to simplify the currently complex pattern of selectivity functions and pattern used in the assessment. A comparative run without time blocks indicates higher Fs for the early period and smaller recruitment pulses though still extra ordinary in the 1960's (Figure 2). In the middle period and in the effective absence of the predominant adult abundance indices (slope and longline survey) the alternate model balances the population by estimating lower SSB than the current model. After 2000 the models are virtually identical.

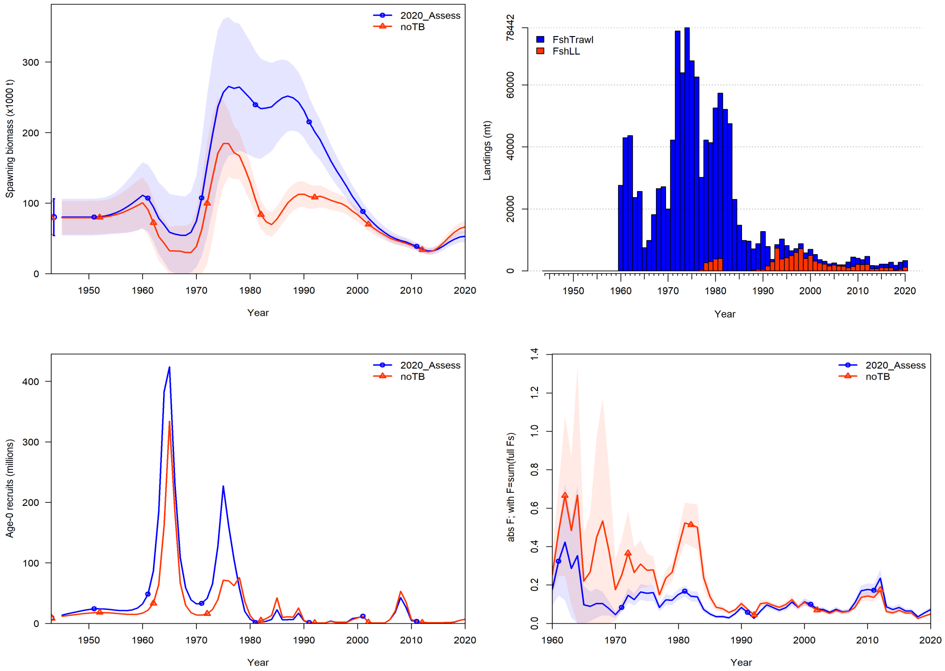


Figure 2. Comparison of current model with an experimental model removing the time blocks on selectivity presented during the meeting.

Therefore, this analysis should be carried out once a selectivity regime has been decided on if those symptoms leading to this thought experiment persist in the model.

Longline survey index:

The longline survey annually alternates between the Bering Sea and the Aleutian Islands with the unsampled proportion in a year being estimated as a proportion of the sampled area. While not ideal as the entire basis of evidence changes each year, it is very important to maintain a better signal on the adult population especially in the absence of the slope survey. The recent poor model fit (post the period used to fix the proportions in the two areas) could be caused by divergence from the original population proportions suggesting the population is not well mixed. This inference is also supported by the admittedly small number of historic tags.

Critically, solely the biomass estimate from the index is used which is problematic as described in the selectivity section. The decision is based on the fact that the length distributions do not match the population trends from the trawl surveys. Invariably conflicting sources of information exist in assessments. Removal of one data source over the other seems subjective to me. It is particularly curious in this situation as the index is retained, its value and importance escribed in the report. The index, however, is equal to the sum of all the lengths, as the proportion of turbot lengths sampled is near 1. Therefore, the lengths are a near random sample of the catch. If the argument is justifiable that they are not a random sample of the survey-vulnerable population as made in this assessment, I fail to see how the biomass index can still be considered a representative sample of the population. It infers that the individuals caught contain more information about the lengths not caught than about their own lengths.

I would recommend an independent review of the LL-survey data to investigate its internal consistency to determine its potential as a biomass index or to reintroduce the length compositions into the penalty function (see also discussion around the cryptic biomass under the section on selectivity function).

Catchability and selectivity:

The choices of catchability and selectivity are key in the way the mode interprets the information and covers a multitude of topics in this assessment. The current model has a somewhat confusing array of time blocks for catchability and selectivity patterns. The reasoning for some of these are rather poorly described in the current assessment report (Meaghan *et al.,* 2020). Going back through older reports one can reconstruct some of the reasoning for specific changes made at a specific time. However, one is left with the impression that changes are largely reactive to assessment issues and the reasons for the choice are rarely reconsidered even when making further adjustments. To develop a more holistic and coherent interpretation of the gear characteristics and the fisheries dynamics taking account of the way these constraints interact in the model.

My considerations on the current selection criteria are:

I have some concerns that the model complexity is near an over-parameterised state. Over parameterisation usually leads to much reduced contrast in the parameters of interest to management with most of the variation in the data being absorbed by the nuisance parameters such as selectivity and catchability. This is often made worse in tier 3 assessments by the coincidental / non-targeted data collection. Contrary, in this assessment the data, particularly the survey data, does seem to contain a high signal to noise ratio as already commented on by the 2007 CIE review despite its coincidental origins.

In its current state, some highly influential parameters in the assessment have been fixed at previous values to aid conversion while others, for which I cannot logically see how the data would contain information on these parameters, are being estimated. Some settings are historic and not re-investigated, while for others the decision is made predominantly on the basis of the log-likelihood where this is heavily influenced by the subjective weighting used in this assessment. Lastly, the criterion does not distinguish between bias and variance. An assessment modelling process that does not exist or is incorrectly formulated is ultimately a poorer management tool than a simpler model that focuses on the dominant known processes known to operate even if the latter has a higher log-likelihood. The reason is that it has better predictability and a more honest assessment of the uncertainties that allow managers to fulfil their objectives reliably.

While there are these general concerns as well as some specific indications of issues discussed below there is no entirely objective categorisation of 'over-parameterised' and I am not suggesting the model fails to fulfil its purpose for management. Rather, I feel the current documentation and approach seem to fall short of providing convincing arguments in favour of high model parsimony. The concern is increased by some rather large residuals and some clear residual patterns in the data.

Time blocks on selectivity:

Although here we have trends in SSB and recruitment but entirely driven by catches. A very flat F does not seem to intuitively justify the complexity of the blocking patterns, i.e., arguments for blocks are that something changed; but other than recruitment, little is suggested to have changed despite some significant anecdotal changes in the management and the fishery. To me this raises the concern that the consistency in F may be a result of the time blocks, acting through the penalty on the F deviates, rather than a reflection of the true dynamics in the fishery. An exploratory analysis conducted at the meeting removed the time blocks from the analysis and resulted in what I would consider a more realistic pattern of historic F given the obvious changes in management. (Figure 2 and previous section).

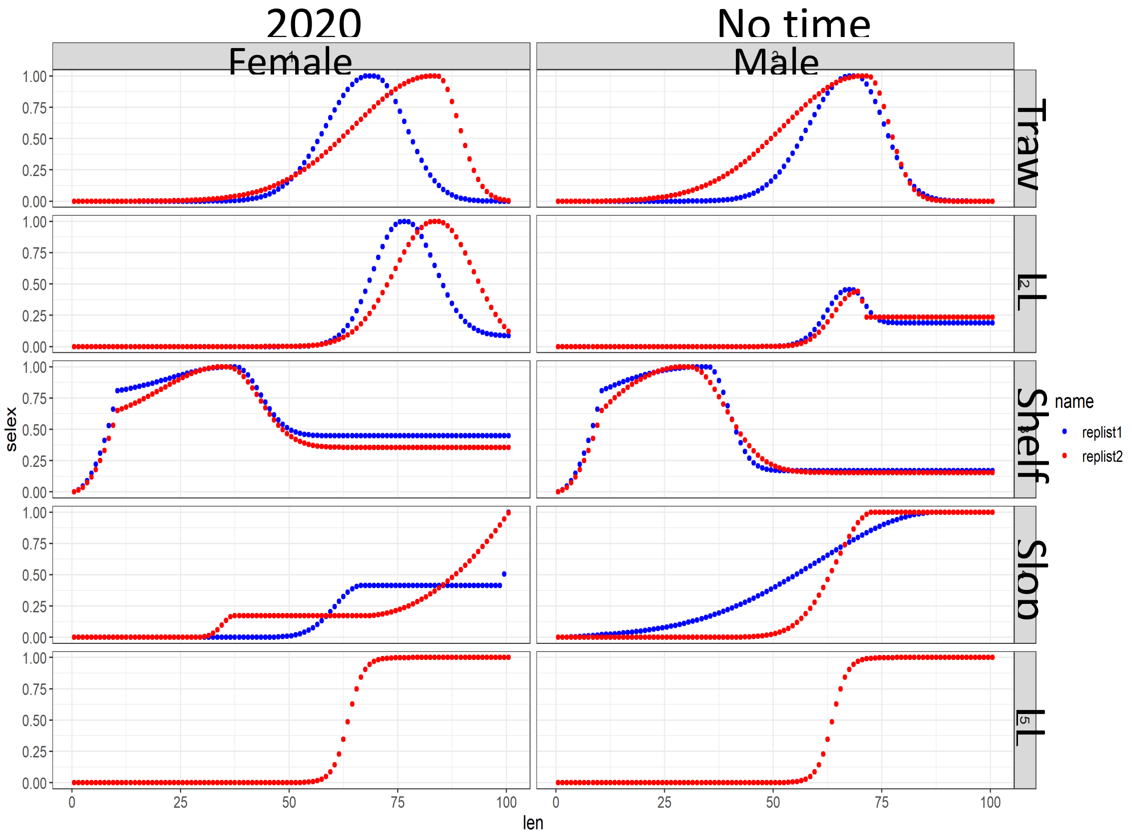


Figure 3: comparisons of final year selectivity patters between the currrent and experimental model removing time blocks.

The argument against this alternate perspective was that it added to the total likelihood, and seemed to ignore the fact that there were only minor differences between the selection patterns between the different time blocks in the original assessment (Bryan *et al.*, 2020) for all but the historic trawl fishery. Other differences were minor, were intertwined with catchability estimates and focused around the terminal selectivities for the dome-shaped functions. In other words, they related to a cryptic biomass element in the stock, suggesting there was little or no evidence in the data to substantiate these differences.

Error! Reference source not found. compares the final year selectivities between the two models. Two things are worth mentioning. Differences in the selectivity curves are predominantly at lengths where there are few observations. The slope survey seems to show some rather dramatic differences in the selectivity, but these are predominantly differences in scaling related to q. Despite the differences there is no easily discernible impact on the most recent stock estimates. The same condition can be explained by fewer parameters in the experimental model suggesting higher parsimony. Having looked at the data in some detail I cannot find any specific trends in the data that would significantly support either of the model interpretation, so one should again conclude that the experimental model has higher parsimony and more anecdotally consistent.

Selectivity functions:

Currently, the selectivities in each time block are estimated as high parameter dome-shaped selectivities for all but the longline survey which is modelled as a logistic function. Differences between sexes within a block and fishery are modelled as off-sets from the females (this is inherent in the SS3 implementation). While the model is nominally consistent with the recommendation to have at least one logistic fleet, the fleet this applies to here is one without any weighted length data in the model. In other words, the purpose of a measure to effectively eliminate the development of a cryptic biomass is circumvented in this model. Any biomass suggested by the other fisheries or surveys is possible simply through a rescaling of catchability. Critically, the other survey catchabilities are fixed, leaving only the critical scaling function of the logistic selectivity to fluctuate.

The age compositional data for the LL survey has been eliminated from the assessment based on the fact that it is not consistent with the recruitment information. While this may be partly justifiable on the basis of the survey design it may be necessary to reinclude this information to attain the benefit of the abundance estimates at older ages. The ghost fits available in the assessment report (Figure 4) persistently indicate the survey should be seeing substantially more adults than are being observed. The deviations appear small on the plots, because they are proportional in relation to the maximum proportion which are driven by the recruitment signals. However, these are in numbers and not in weight, and most of the catches represent immatures. If the survey selectivity is indeed logistical, then the biomass estimates from the current model are likely to be optimistic, and recent trends in SSB underrepresent the true fluctuations in biomass.

An attempt was made at the review to estimate the cryptic biomass, but there was some uncertainty around what the numbers supplied by SS3 were and the proportion of the spawning biomass that were evading exploitation were either too high or too low. With hindsight, it seems somewhat irrelevant to calculate the proportion of the exploitable biomass. What is relevant is the unobserved biomass (i.e., including surveys), which in this assessment seems to be potentially large under certain stock conditions.

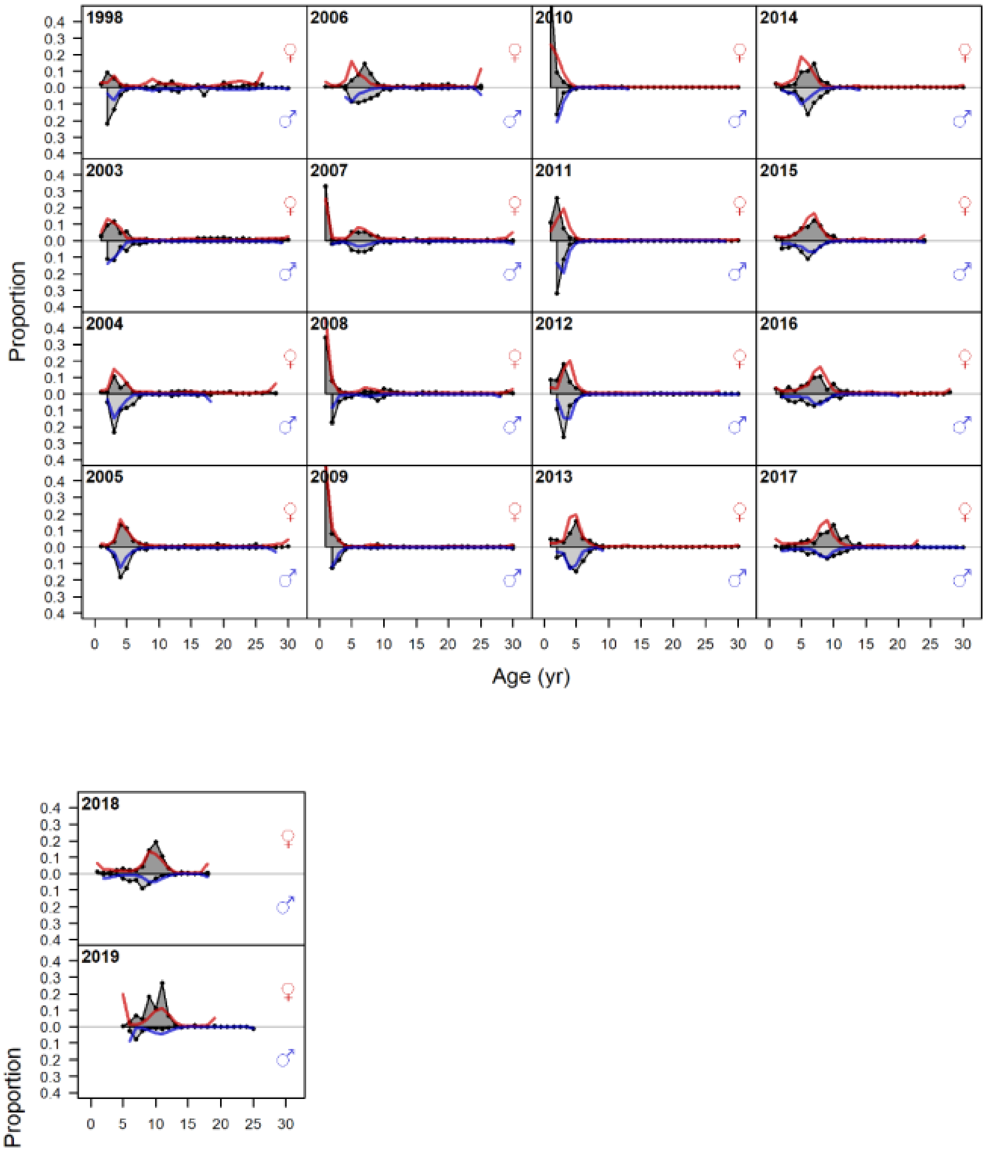


Figure 4: Model 16.4a (2020) shelf survey age composition data and "ghost" fits (red and blue line "Ghost" fits are projected fits as they are not fit to the likelihood for the age composition

Sex off-sets for selectivities:

The assessment models selectivities separate by sex through the offset function. From a gear selectivity perspective there appears to be little evidence to suggest this is necessary as length-selectivity is estimated, and the gears seem to select predominantly on the basis of length. One issue that does arise is that turbot migrate from the shelf nursery areas to the slope with maturity and the maturity for the sexes occurs at different lengths due to the growth dimorphism. When implementing selectivities using the gender off-set, that risk has to be weighed up against the risk of trying to estimate inestimable parameters in strongly dimorphic species. It is for example not clear to me how the model can estimate the size of the off-set in a parameter determining the selectivities in male turbot of over 90cm, when males do not grow to this size. Generally, parameters without information reduce model stability (see subsequent section) and often parameters serving different purposes / processes have to be fixed in order to make them estimable.

The appropriate solution would be to estimate the selectivities entirely independently, but this is currently not available in SS3. The justification of the off-set function is to save parameters/increase parsimony. However, given the generosity with which this particular model deals with parameters, I would strongly recommend that this would be a better solution to issue in this assessment. Removing a single hard to justify time block presents a much greater saving in parameters than those saved by the offset. At the very least being able to do so to examine the impact as a diagnostic tool would be a very useful addition to the SS3 frame work.

Survey selectivities and catchabilities:

This assessment has taken the somewhat unusual route for implementing time blocks in trawl surveys. Surveys is what we aim to keep as consistent as possible so adding time blocks is only really justifiable as a last resort and when there is specific process knowledge as a justification. For example, the reduction in two durations in the shelf survey may proportionally affect the larger individuals more than the smaller ones as the former may be able to maintain higher swimming speeds for a short duration, i.e., their selectivity would decrease with a decrease in duration. Density dependent effects may disperse the individuals to an area outside the survey so that the proportion sampled is reduced.

During the discussion, the argument made for the selectivity time blocks was that it was necessary to deal with specific large cohorts which were destabilising the model. To me this seems like muting the very signal you are interested in, so this should be reconsidered. The systematic nature of the residual patterns from the current and experimental model show that the likelihood improvement of the time blocks comes predominantly from down scaling a few very large residuals that still remain large even in the current model. Moreover, the strong year classes which it was suggested the blocking was aimed to deal with is still substantially underestimated given the length data (**Error! Reference source not found.**). They are equally visible in the other compositional information also so are either coming up against parameter bounds, their growth is poorly modelled or they are simply inconsistent with the abundance information. Introduction of time blocks does not seem to serve the intended purpose and is at best detracting from some other process issues in the assessment.

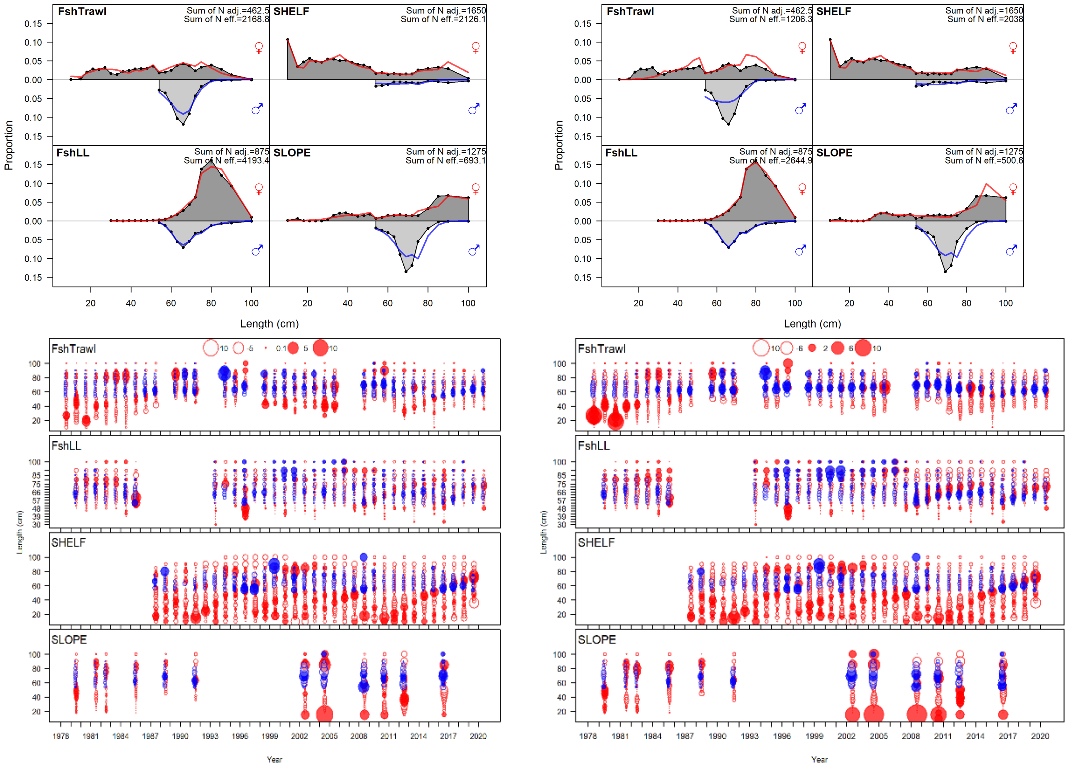


Figure 5: Comparison of length residuals between the current model and an experimental model without time blocks, showing that the cohort pattern is still equally discernible for the surveys and a relatively small number of samples disproportionately contributes to the improvement in the log-likelihood. The exception appears to be the early time block for the trawl fishery (pre-1985).

N.B In the above plots it is apparent that strong cohorts are underestimated in all compositional data sources, i.e., they appear not balance each other out. Other model constraints such as the seemingly high choice of sigma-R (fixed at 0.6) for a stock that has very sporadic large recruitment may contribute to this effect and complicate the appropriate parameterisation of the selectivities.

I would suggest proceeding as follows:

Document the reasons and timings for changes in selection, followed by a thorough review of their likely interactions and implications for management.

- Based on the previous findings, a prioritisation of the selection should allow for a significant simplification of the selectivities and an improved model parsimony.
- From this basic model, one can then move forward to stepwise increase complexity, where deemed appropriate, based on process, model precision and management considerations.
- An alternative would be to examine the parameter stability from the current assessment and remove or further constrain parameters based on the jitter and MCMC analyses. This would necessarily be a sequential ‘one-out’ analysis and tends to constrain the thinking to the realm of the current implementation options. Additionally, constraints such as time blocks and notably their interaction with the parameters is not easily considered in this approach. (See further details under model stability.)

Other discussions:

Fixing survey catchabilities:

In the current assessment catchabilities are fixed for the trawl surveys. When assessments are interpreted predominantly in a relative sense such as tier 4 assessments which take account of this fact in their more precautionary management approach. Catchabilities are important scaling parameters and become more relevant to management as stocks aim to progress towards higher tiers. In all cases it is necessary to consider why such quantities are not easily estimated when deciding on approaches to fix them, especially in this case where the value apparently was estimable at a previous point in the assessment.

In this case, I suspect the removal of the compositional data from the long-line survey has contributed significantly to this uncertainty. I have already discussed concerns regarding the removal of this data in relation to the cryptic biomass issue (previous section), but generally believe the choice of using previously estimated values to be largely subjective. Additionally, this may contribute to considerable instability in the estimate of the LL-survey as illustrated in the Table 5.19 in the assessment report. I would hope the proposed review of the selectivities will deal with this issue at the same time and resolve the issue.

Skip spawning:

During the assessment presentation some research to investigate skip spawning in the species was proposed. This represents important biological research in understanding the ecology and biology of the species. For management of this species the results are unlikely to influential unless significant variation in skip spawning over time can be demonstrated and more importantly predicted for future time periods. The reason for this is that if skip spawning exists it will lower the estimate of SSB the entire time period and equally effect affect virgin SSB so that the ratio, i.e., the effective management quantity for tier 3a will remain the same. If steepness was estimated in the assessment there might be a rescaling in the parameter, and with it fixed, there may be some slight differences in the estimation of mean recruitment at very low stock levels.

Retrospective analysis:

Bryan *et al.* (2020) conclude that there is retrospective bias in the estimation of the 2009 cohort with subsequent impacts on SSB as the cohort matures. They conclude:

> *Therefore, there is some uncertainty about the adult portion of this stock on the slope. Uncertainty in assessment model results due to missing the most recent EBS shelf bottom trawl survey was evaluated in Bryan et al. (2020). They found that the direction and magnitude of retrospective bias was an important determinant in the level of expected uncertainty in our stock assessment results.*

While I agree that the lack of the EBS slope survey is a significant contributor to the increasing uncertainty regarding stock status, I find it difficult to justify it on the basis of the retrospective analysis in this assessment. Although the peels in the retrospective are produced at an annual time step, they do not account for the data availability. For the slope survey they are only two data points in the retrospective. Removal of the 2016 (2015 peel) data point results in the marked decrease in the 2009-recruitment estimate. Removal of the 2012 data (2011 peel) shows no further change beyond the annual background bias estimate. Though it is not a fair retrospective comparison since the selectivity time block at this stage contains only one data point Therefore the predominant part of the change in the estimate of the 2009 cohort is difficult to classify as a retrospective bias.

The smaller underlying retrospective bias seems to be caused by a persistent difference in trend between the shelf survey and one of the other data sources, catch or LL-survey. I suspect the issue is in the catch compositional data as already explained previously the LL-survey is largely ineffective at describing recruitments.

Model stability (MCMC and jitter):

In general, I consider the approaches taken in this assessment to be suitable to assure robustness of advice. The information presented specifically identified some weaknesses in the model. However, the presentation of the results and the length of the MCMC chains left questions unanswered which would have to be considered before applying the assessment to advice. There was insufficient time at the meeting for the assessor to conduct the necessary analysis to address these concerns fully.

The MCMC analysis indicated that the model was having difficulty uniquely identifying some parameters. While the output metrics seemed to be reasonably stable different combinations of parameter appeared to be possible to arrive at those metrics. The jitter analysis suggested the gradient was not particularly uniform, and the model did manage to find a number of different solutions from different starting points. The biggest issue seemed to be around the estimation of the catchability and selectivity parameters confirming some of the issues previously described on the basis of process considerations.

An interesting discussion and examination of specific parameters and potential model modifications was conducted. While I learned a lot about the general interpretation of results from both methods, my main take from the discussion was that it confirmed the concerns of model parsimony and structural robustness. In my opinion specific interpretation of the results were not possible, because the MCMC chains were likely to short but mostly because it had already been suggested that the selectivity regime was to be reviewed at which point the current results will be uninformative.

Whale depredation in LL-survey:

During the presentation of the LL-survey specific mention was made of the efforts to identify samples which were likely unrepresentative of the catches due to the removal of individuals from the sampling gear during retrieval by whales. The data presented made a convincing argument that staff were able to reasonably and reliably identify affected sets with these data being removed from CPUE analysis. Time series and spatial information on the proportion of sets being affected suggest that the number can at times be large. However, there is little trend in the proportion in the Bering Sea, and only a slight indication of an increase in the Aleutian Islands part of the survey.

While I agree that the benefits of including the data outweigh the risks of having no fisheries independent means of monitoring the adult population in the absence of the slope survey, possibly more could be done to investigate this aspect further while also considering how in future are more effective sampling protocol could be developed.

Considerations:

- Whales are unlikely to target sets randomly. Those with higher catches are likely to be more prone to interference than those with low or 0 catch. This potentially results in a biased estimate of CPUE even if the targeted sets are excluded.

- The long-line fishery has declined quite dramatically and during the discussions with the survey and observer programs there were anecdotal comments on an increasing pressure on this fishery due to whale depredation which may be associated with its decline. This contradicts the results for the survey which indicate at best a minor increase in the level of depredation. Either an increase in the whale population or a dissemination of behaviours to exploit fishing gears could be responsible for effects in the survey index.

Summary of findings Gulf of Alaska Northern and Southern Rock Sole assessment

The current statistical catch-at-age assessment methodology implemented in SS3 was developed in 2017 (Bryan *et al.*, 2017) with a small change from the previous full assessment conducted in 2015. The stock is now on a four-year assessment cycle and this meeting is to provide a review and possible recommendations for the 2021 assessments. The stock is currently managed as part of the 'shallow-water flatfish complex' under Tier 3a rules having commendably developed historically from Tier 4 biomass index-based methods.

On initial review of the assessments (northern and southern rock sole), I had considerable concerns regarding its ability to follow the trends in the stock dynamics with the model seemingly over-smoothing much like a biomass dynamic model through the middle of the data and suggesting little trend while not fitting the annual length comps from survey and fishery particularly well. It is important to note that the species/stocks are very similar in their ecology and the assessments are near identical including the data sources used. They are also linked in the sense that historically catches were only recorded as 'rock soles'.

While the initial concerning symptoms remain, I was able to gain confidence through a more detailed examination of the survey data (estimated age raised population number from the survey provided by the assessor) that the data themselves suggested that there had been a relatively little trend in the dynamics over the survey period and that there were some inherent year-effects in the GOA trawl survey, similar to those interpreted by the model. The smooth trend therefore came from the data themselves and was not a result of conflicting information from different data sources. The relatively poor fit to the compositional data is most likely a less than ideal specification of growth and or an interaction with the way selectivities are specified. There is little doubt in my mind that even if the trends in SSB and F were to change somewhat through model improvements, the scaling of the management metrics are robust for management under tier 3 a consideration. Given the low exploitation rate and the associated quota that is not currently restricting the fishing activity, this fishery seems to largely manage itself well within sustainable limits. Under current circumstances, I consider the model well able to fulfil the purpose of management and consider the recommendations made herein mainly forward looking/exploratory to ensure that if and when the situation changes the appropriate evidence base exists to support management.

This should not be interpreted as a suggestion that this is not the best scientific evidence available for the assessment process. Some of the diagnostics and theories suggest alternate approaches may be beneficial for some data sources, but there is currently no evidence to deliver a verdict to say that they would be substantially different let alone better. On current evidence, I do believe that this is the best scientific evidence base on which to provide the 2022 advice.

Prioritization of research:

I prioritised the assessment needs as follows:

1) Improvements in modelling growth are necessary to attain better fits to the length distributions in both assessments, though it is more important for the Northern rock sole assessment. The review was able to identify some potential causes based on the data, but a resolution requires further work, both to determine the actual cause and the development of models to facilitate such hypothesis. An appropriate interim measure may be to move to age-based selection using the internally consistent survey age information as the basis.
2) Selectivities have comparatively few parameters in this model when compared to other models. Despite this there is still some tendency for parameters not to be uniquely identifiable. This could be due to the poor modelling of growth, but the fishery represents a single fleet using trawls within a confined space; so, it is not clear to me why a dome-shaped selection is 'likely' as stated in the assessment report, especially since the survey uses a similar gear and is modelled as a monotonic function. Some efforts to simplify selectivity may still be necessary after growth is modelled more appropriately and definitely required if the latter cannot be adequately resolved.
3) The survey data for Northern rock sole show some internal inconsistencies with regards to year effects. The model is currently picking those up and interpreting them as residuals so it is not a big issue. However, it may be advantageous to develop a model-based index for this species that might then potentially account / explain why these effects occurred. This would reduce overall model uncertainty and potentially aid convergence further.

Review activities Gulf of Alaska Northern and Southern Rock Sole assessment

Catch data:

Historic catch is split 50:50 between the two species in this assessment. More recent landings data are split according to the proportions of the two species in the observer program. There were some concerns raised in the review regarding the species split for two reasons. The survey data have a split closer to 70:30 and the more recent observer data suggest that there is considerable interannual variation in the proportions.

Taken together, the two largely independent assessments resolve the difference in catch ratio between survey and fishery reasonably consistently through differences in selectivity. This is not to say that those choices of selectivity are not forced by the catches used in the assessment, but one would expect to notice a visible change in the retrospective pattern between the period where 50:50 ratio was used and the more recent period where observed ratios are used.

During discussion the option of treating the reported catch as less precise would be useful to check if there is negative correlation between the catch residuals between the two species as an indication the need to find alternative remedies for assigning catches to species. Personally, I do not consider this investigation as high priority mainly because of the fixing of survey catchability in both assessments. This measure essentially scales the populations and somewhat surprisingly for two different species they share considerable similarity in their recruitment dynamics so that large changes in the proportion of catches are relatively unlikely and, in any case, difficult to estimate, particularly considering the low exploitation rate suggested.

Growth:

The appropriate specification of growth is important in models not using conditional age-composition data. There are obvious benefits to the approach including being able to specify selection at length rather than age but there are also inherent difficulties in appropriately specifying selection when growth models are uncertain or imprecise. Unfortunately, in the case of rock sole, particularly the northern rock soles the model struggles to accommodate some large systematic length residuals. For males the model seems to estimate survey size-at-age correctly but fails to accommodate the variance properly, while for females both the modal length and the variance appear to be misspecified. The fishery data is more appropriately fitted overall, but at the annual level suffers from systematic variation which could be attributable to high sample autocorrelation or constraint in estimating variation in cohort strength.

Sampling design:

The specification of growth in the rock sole assessment like the turbot assessment suffers from a misspecification of growth because the existing age subsampling process of the GOA survey are not appropriately accommodated to reflect the variation in the proportion of ages sampled in the different length categories. The data is treated as random samples for all years though only appropriate for the last year of survey data where the subsampling design was altered.

Taking full account of the statistical properties of the data in SS3 is currently difficult because length and age compositions are treated independently in terms of their weights with the effective sample size being attributed to the whole length sample not specifically by length. In addition, the various approaches to model weighting will inevitably lead to departure from the statistical theoretical weighting in the sampling design due to the estimation of the effective sample size. Therefore, a correct interpretation of the length stratified data is currently difficult in SS3 something the authors clearly already struggled with as indicated by the number of models presented in the report differing only in the data weighting approaches. I recommend examining the possible magnitude of the impact on the assessment would be to compare the differences in the estimated size-at-age from the model with the design-based estimates obtainable from the survey database to check for persistent biases at age and by cohort. If significant differences are apparent then it may be preferential to use the design-based mean size-at-age from the survey directly as recommended for the turbot assessment.

Variation in growth:

Variation in growth when specifying a constant growth curve within the models generally tends to overly smooth the recruitment deviates. This is particularly problematic when the estimate of growth falls between two periods of differing growth, i.e., is not appropriate for any cohorts and leads to the sorts of residuals observed for the survey in this model. However, closer examination of the data does not seem to suggest a substantial difference in the growth rates over time. What is suggested by the data is a bifurcation of the size-at-age within each year particularly in the northern rock sole data (Figure 6).

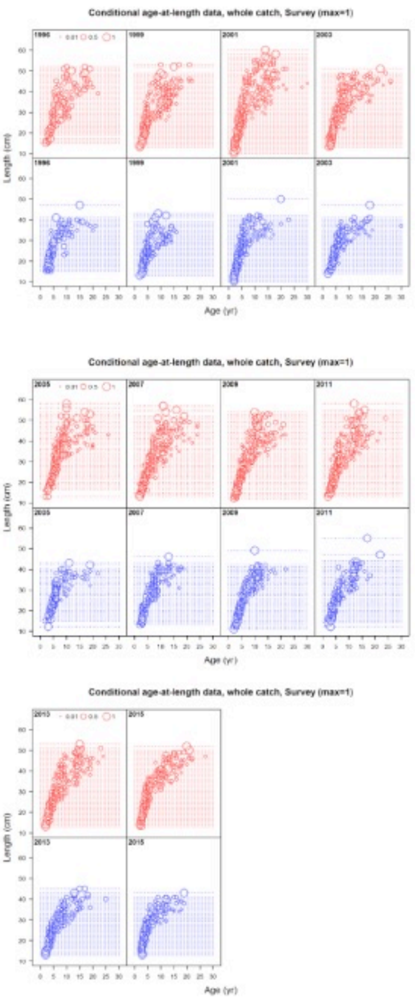


Figure 6: Northern rock sole conditional age-at-length data indicating a divergent growth pattern for both sexes after age 5 -6 corresponding to the onset of maturity.

Aging errors are unlikely to lead to such bimodal data particularly since aging estimates are thought to be particularly precise for this species. Several alternate plausible explanations for this phenomenon were discussed in during the review:

- Species identification criteria may not be as accurate as thought and the northern rock sole represent a mix of two different species. However, the issue although less obvious is also present in southern rock sole where all fish seem to grow as fast of faster than the faster growing individuals identified as northern rock sole, i.e., there would need to be three species.
- Environmentally induced spatial differences in growth could result in the observed pattern of variation. The area around Kodiak Island is identified as prime habitat for both species proportionally

differentiated in distribution by depth. Environmental or ecological processes could lead to different growth rates in both species at different depth but we would tend to expect the divergence of growth to occur at all ages. Interregional differences seem a more likely scenario although still not clear why the differences are more pronounced after age 5/6.

- Intra species life history strategies could persist in what seems in all likelihood a relatively recent speciation event. The premise of skip spawning is that the individual maximises its reproductive potential by redirecting reproductive output in a given year to growth in order to harvest the benefits of greater fecundity with greater size in future years. However, the differences between the two growth curves appear much greater than the annual growth increments. Also, we would expect that different individuals would skip spawn in different years so that after a number of years size-at-age should converge towards a single size, i.e., have the same Lmax but different k. Delay of maturation is more likely to produce the observed effect unless growth compensation is large.
- Relative genetic isolation within the stock be it spatial or reproductive resulting in different growth to me still seem the most likely scenario, but immigration from other populations to the south and north could produce similar effects and would probably best explain the divergence only at the onset of maturity when flatfish tend to increase their movements when reaching maturity.

Unfortunately, there was little time to make progress on the identifying the cause of the growth differences. Because of the lack of understanding it was not possible to develop remedial measures to account for the observed age information. Nor was it possible to judge the likely magnitude of the effect, since the data were provided only as the proportion of ages at length. This may be a high proportion of the individuals at a specific length but it may still be an insignificant proportion of the total sample. However, the poor fit to the length data suggests it is important and will ultimately interact with the appropriate specification of selectivity (see later section).

Surveys:

The assessments rely on a single source of fisheries independent observations from the GOA trawl survey which are treated as an absolute abundance estimate with catchability fixed at 1. These facts make the survey highly influential in the assessment, so care must be taken to ensure correct interpretation of the data. Appropriately then, this was an important topic at the meeting and this data and associated risks were discussed in detail.

The fixing of catchability at 1 is of concern. The biomass index is raised to the area and provides a population estimate. However, there are untrawlable areas which cannot be effectively sampled with this gear. The concern is that these areas may harbour a significant proportion of the population. While I commend the efforts of the AFSC center to explore methods of estimating abundance in these areas, generally I feel there is considerably more need for this for other species known to prefer rocky or rough habitats which were the dominant reason given for untrawlability.

- Proportion of population in untrawlable habitats is likely to be relatively small due to habitat preferences and relatively small proportion of untrawlable grounds.

- The proportion of habitat that cannot be sampled is constant, and at the currently low levels of exploitation one would not expect the proportion of the population between the two habitats to change significantly over time at comparatively low exploitation levels. The management for these species is based on tier 3a rules which are relative metrics of stock status and therefore are unlikely to be influenced by adding a fixed proportion to the population.
- Lastly, the assessments lack the expected retrospective patterns in F and SSB that one would associate with an effectively unclosed population monitored by the survey.

Internal consistency:

Given my concerns regarding the lack of contrast in the assessment I requested to look at the survey data independently. The assessor was able to provide the data in the format of age-disaggregated index from which I plotted catch curves and checked for cohort consistency using means standardised (to remove selectivity-at-age effects) indices of abundance-at-age.

The catch curves suggested Z was around 0.2 for Northern rock sole and 0.3 for Southern rock sole with some interannual variability but with no discernible trend between cohorts, i.e., roughly constant Z over the time of the survey (Figure 7). The assessments indicate Z to be slightly higher with M fixed at levels close to the indicated Z from the survey analysis, but otherwise largely consistent with respect to the relative scaling between the two species and the lack of a clear trend in F. Estimated variability in cohort strength is slightly larger by the survey but the modelling process is expected to smooth data somewhat.

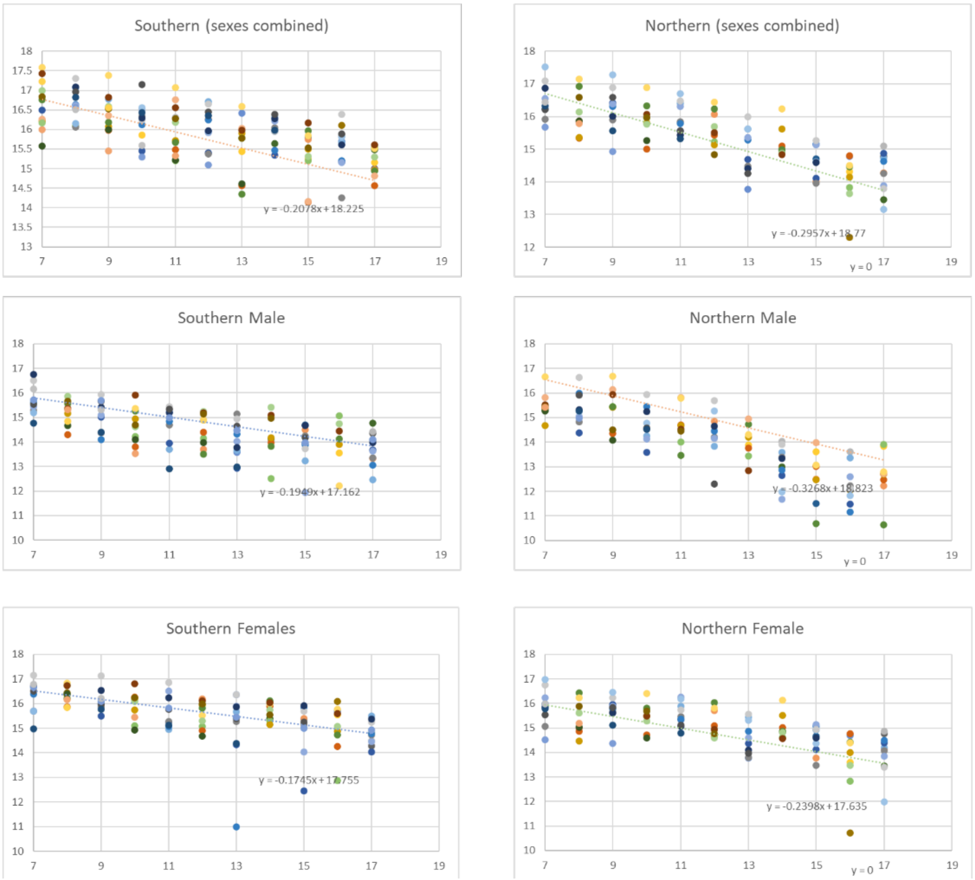
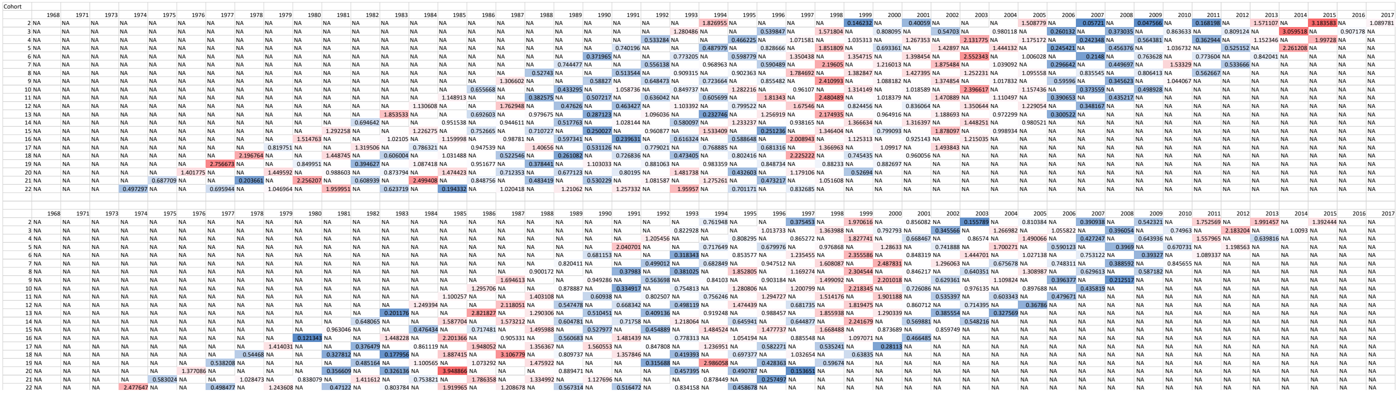


Figure 7: Catch curves, assuming full selection at age 7 for Southern (left) and Northern (right) rock sole. Combined sexes and males and females separately. For each figure one cohort provides the slope (formula) just for reference. Colours represent cohorts and the alternating colour banding at odd and even ages results from the biannual survey periodicity.

The means standardised catch-at-age tables (Figure 8) demonstrate a generally high internal consistency in the proportion of the mean catch-at-age observed at different ages for each cohort (columns). Striking in this is that periods of low and high recruitment show some synchrony between the two species although within the period different cohorts are indicated to be the strongest for each species. This suggests there are some common large-scale environmental or ecological effects driving recruitment synchronously in both species with some stochasticity overlayed. Lastly, some year/survey effects in the survey are apparent particularly for northern rock sole, as the last two surveys indicate that for most cohorts they suggest an abrupt and unexpected decrease in the proportion of the mean abundance for these cohorts, while the 2009 and 2011 surveys indicate generally higher proportions. These internal inconsistencies are therefore entirely consistent with the interpretation from the northern rock sole assessment. For southern rock sole, the assessment generally fits the survey biomass index more closely and the survey data indicate an even greater internal consistency for this species. The exception is the 2011 survey data which the assessment underestimates while there is no clear indication of a negative year effect from the survey data themselves.



Cohort	1968	1971	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.826955	NA	NA	NA	NA	0.146232	NA	0.40059	NA	NA	NA	1.508779	NA	0.05721	NA	0.047566	NA	0.168198	NA	1.571107	NA	3.183583	NA	1.089781
3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.280486	NA	NA	0.539847	NA	1.571804	NA	0.808095	NA	0.54703	NA	0.980118	NA	0.260132	NA	0.373035	NA	0.863633	NA	0.809124	NA	3.059518	NA	0.907178	NA
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.533284	NA	NA	0.466225	NA	1.071581	NA	1.035313	NA	1.267353	NA	2.131775	NA	1.175172	NA	0.242348	NA	0.564381	NA	0.362944	NA	1.152346	NA	1.99728	NA	NA
5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.740196	NA	NA	0.487979	NA	0.828666	NA	1.851809	NA	0.693361	NA	1.42897	NA	1.444132	NA	0.245421	NA	0.456376	NA	1.036732	NA	0.525152	NA	2.261208	NA	NA	NA
6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.371965	NA	NA	0.773205	NA	0.598779	NA	1.350438	NA	1.354715	NA	1.398454	NA	2.552343	NA	1.006028	NA	0.2148	NA	0.763628	NA	0.773604	NA	0.842041	NA	NA	NA	NA
7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.744477	NA	NA	0.556138	NA	0.968963	NA	0.590489	NA	2.19605	NA	1.216013	NA	1.875484	NA	1.039092	NA	0.296642	NA	0.449697	NA	1.53329	NA	0.533666	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.52743	NA	NA	0.513544	NA	0.909315	NA	0.902363	NA	1.784692	NA	1.382847	NA	1.427395	NA	1.252231	NA	1.095558	NA	0.835545	NA	0.806413	NA	0.562667	NA	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.306602	NA	NA	0.58827	NA	0.648473	NA	0.723664	NA	0.855482	NA	2.410993	NA	1.088182	NA	1.374854	NA	1.017832	NA	0.59596	NA	0.345623	NA	1.044067	NA	NA	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.655668	NA	NA	0.433295	NA	1.058736	NA	0.849737	NA	1.282216	NA	0.96107	NA	1.314149	NA	1.018589	NA	2.396617	NA	1.157436	NA	0.373559	NA	0.498928	NA	NA	NA	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.148913	NA	NA	0.382575	NA	0.507217	NA	0.636042	NA	0.605699	NA	1.81343	NA	2.480489	NA	1.018379	NA	1.470889	NA	1.110497	NA	0.390653	NA	0.435217	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.130608	NA	NA	1.762948	NA	0.47626	NA	0.463427	NA	1.103392	NA	0.799522	NA	1.67546	NA	0.824456	NA	0.836064	NA	1.350644	NA	1.229054	NA	0.348167	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.853533	NA	NA	0.692603	NA	0.979675	NA	0.287123	NA	1.096036	NA	0.232746	NA	1.256919	NA	2.174935	NA	0.964916	NA	1.188693	NA	0.972299	NA	0.300522	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.694642	NA	NA	0.951538	NA	0.944611	NA	0.517763	NA	1.028144	NA	0.580097	NA	1.233237	NA	0.938165	NA	1.366634	NA	1.316397	NA	1.448251	NA	0.980521	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.292258	NA	NA	1.226275	NA	0.752665	NA	0.710727	NA	0.250027	NA	0.960877	NA	1.533409	NA	0.251236	NA	1.346404	NA	0.799093	NA	1.878097	NA	0.998934	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
16	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.514763	NA	NA	1.02105	NA	1.159998	NA	0.98781	NA	0.597341	NA	0.239631	NA	0.616324	NA	0.588648	NA	2.008943	NA	1.125313	NA	0.925143	NA	1.215035	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	NA	NA	NA	NA	NA	NA	NA	NA	0.819751	NA	NA	1.319506	NA	0.786321	NA	0.947539	NA	1.40656	NA	0.531126	NA	0.779021	NA	0.768885	NA	0.681316	NA	1.366963	NA	1.09917	NA	1.493843	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	NA	NA	NA	NA	NA	NA	NA	2.196764	NA	NA	1.448745	NA	0.606004	NA	1.031488	NA	0.522546	NA	0.261082	NA	0.726836	NA	0.473405	NA	0.802416	NA	2.225222	NA	0.745435	NA	0.960056	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	NA	NA	NA	NA	NA	NA	2.756673	NA	NA	0.849951	NA	0.394627	NA	1.087418	NA	0.951677	NA	0.378441	NA	1.103033	NA	0.881063	NA	0.983359	NA	0.848734	NA	0.88233	NA	0.882697	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	NA	NA	NA	NA	NA	1.401775	NA	NA	1.449592	NA	0.988603	NA	0.873794	NA	1.474423	NA	0.712353	NA	0.677123	NA	0.80195	NA	1.481738	NA	0.432603	NA	1.179106	NA	0.52694	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	NA	NA	NA	NA	0.687709	NA	NA	0.203661	NA	2.256207	NA	0.608939	NA	2.499408	NA	0.848756	NA	0.483419	NA	0.530229	NA	1.081587	NA	1.275261	NA	0.473217	NA	1.051608	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	NA	NA	NA	0.497297	NA	NA	0.695944	NA	1.046964	NA	1.959951	NA	0.623719	NA	0.194332	NA	1.020418	NA	1.21062	NA	1.257332	NA	1.95957	NA	0.701171	NA	0.832685	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

	1968	1971	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.761948	NA	NA	0.375453	NA	1.970616	NA	0.856082	NA	0.155789	NA	0.810384	NA	0.390938	NA	0.542321	NA	1.752569	NA	1.991457	NA	1.392444	NA	NA
3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.822928	NA	NA	1.013733	NA	1.363988	NA	0.792793	NA	0.345566	NA	1.266982	NA	1.055822	NA	0.396054	NA	0.74963	NA	2.183204	NA	1.0093	NA	NA	NA
4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.205456	NA	NA	0.808295	NA	0.865272	NA	1.827741	NA	0.668467	NA	0.86574	NA	1.490066	NA	0.427247	NA	0.643936	NA	1.557965	NA	0.639816	NA	NA	NA	NA
5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2.040701	NA	NA	0.717649	NA	0.679976	NA	0.976868	NA	1.28633	NA	0.741888	NA	1.700271	NA	0.590123	NA	0.3969	NA	0.670731	NA	1.198563	NA	NA	NA	NA	NA
6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.681153	NA	NA	0.318343	NA	0.853577	NA	1.235455	NA	2.355586	NA	0.848319	NA	1.444701	NA	1.027138	NA	0.753122	NA	0.39327	NA	1.089337	NA	NA	NA	NA	NA	NA
7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.820411	NA	NA	0.499012	NA	0.682849	NA	0.947512	NA	1.608087	NA	2.487831	NA	1.296063	NA	0.675678	NA	0.748311	NA	0.388592	NA	0.845655	NA	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.900172	NA	NA	0.37983	NA	0.381025	NA	1.852805	NA	1.169274	NA	2.304544	NA	0.846217	NA	0.640351	NA	1.308987	NA	0.629613	NA	0.587182	NA	NA	NA	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.694613	NA	NA	0.949286	NA	0.563698	NA	0.84103	NA	0.903184	NA	1.499092	NA	2.201018	NA	0.629361	NA	1.109824	NA	0.396377	NA	0.212517	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.295706	NA	NA	0.878887	NA	0.334917	NA	0.754813	NA	1.280806	NA	1.200799	NA	2.218345	NA	0.726086	NA	0.976135	NA	0.897688	NA	0.435819	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.100257	NA	NA	1.403108	NA	0.60938	NA	0.802507	NA	0.756246	NA	1.294727	NA	1.514176	NA	1.901188	NA	0.535397	NA	0.603343	NA	0.479671	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.249394	NA	NA	2.118051	NA	0.547478	NA	0.668342	NA	0.498119	NA	1.474439	NA	0.681735	NA	1.819475	NA	0.860712	NA	0.714395	NA	0.36786	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.201176	NA	NA	2.821827	NA	1.290306	NA	0.510451	NA	0.409136	NA	0.919248	NA	0.988457	NA	1.855938	NA	1.290339	NA	0.385554	NA	0.327569	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.648065	NA	NA	1.587704	NA	1.573212	NA	0.604781	NA	0.71758	NA	1.218064	NA	0.645941	NA	0.644877	NA	2.241679	NA	0.569881	NA	0.548216	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.963046	NA	NA	0.476434	NA	0.717481	NA	1.495988	NA	0.527977	NA	0.454889	NA	1.484524	NA	1.477737	NA	1.668488	NA	0.873689	NA	0.859749	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
16	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.121343	NA	NA	1.448228	NA	2.201366	NA	0.905331	NA	0.560683	NA	1.481439	NA	0.778313	NA	1.054194	NA	0.885548	NA	1.097071	NA	0.466485	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	NA	NA	NA	NA	NA	NA	NA	NA	1.414031	NA	NA	0.376479	NA	0.861119	NA	1.948052	NA	1.356367	NA	1.560553	NA	0.847808	NA	1.236951	NA	0.582271	NA	0.535241	NA	0.28113	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	NA	NA	NA	NA	NA	NA	NA	0.54468	NA	NA	0.327812	NA	0.177956	NA	1.887415	NA	3.106779	NA	0.809737	NA	1.357846	NA	0.419393	NA	0.697377	NA	1.032654	NA	0.63835	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	NA	NA	NA	NA	NA	NA	0.538208	NA	NA	NA	NA	0.485164	NA	1.100565	NA	1.073292	NA	1.475922	NA	NA	NA	0.315688	NA	2.986058	NA	0.428363	NA	0.59674	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	NA	NA	NA	NA	NA	1.377086	NA	NA	NA	NA	0.356609	NA	0.326136	NA	3.948866	NA	NA	NA	0.889471	NA	NA	NA	0.457395	NA	0.490787	NA	0.153651	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
21	NA	NA	NA	NA	0.583024	NA	NA	1.028473	NA	0.838079	NA	1.411612	NA	0.753821	NA	1.786358	NA	1.334992	NA	1.127696	NA	NA	NA	0.878449	NA	0.257497	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	NA	NA	NA	2.477647	NA	NA	0.498477	NA	1.243608	NA	0.47122	NA	0.803784	NA	1.919965	NA	1.208678	NA	0.567314	NA	0.516472	NA	0.834158	NA	0.458678	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Figure 8: Means standardised abundance for the GOA trawl survey for Southern (top) and Northern (bottom) rock sole sexes combined. Numbers-at-age/ average Numbers-at-age conditionally formatted with red indicating above average and blue below average. Columns represent cohorts and rows ages 2-22. Diagonal lines (top right to bottom left) each represent a survey in alternating years.

My conclusions from this analysis are:

- Irrespective of the apparent lack of contrast in the model output and the concerns over model parsimony, the assessment is picking up these dynamics directly from what appears to be high quality survey data rather than just smoothing through a cloud of data points.
- Somewhat surprisingly the survey shows stronger year effects for Northern rock sole, which is more widely distributed. Usually, the more contiguously distributed species suffer from generally (i.e., strata not specifically chosen for the species) stratified random surveys. Here it may be simply that the strata are better suited for the southern than the northern rock sole.
- Northern rock sole may benefit from a model-based survey index to account for the changes in survey effort and the interannual variation in survey station placement.
- I understand the desire to use age compositional data marginally in general, but when the information trend in the compositional data is this consistent and avoids the issue of trying to model growth appropriately, I think it is worth considering reintroducing the age compositions more formally even if at this point it is not possible to appropriately accommodate the length stratified otolith sampling regime in SS3.

Selectivity:

The implementation of selectivities in the rock sole models are comparatively simple compared to the turbot assessment as constant selectivity is assumed for both fishery and survey. The most consistent compositional data from the survey is modelled as an asymptotic function while the current model and the proposed modified model allow for a decrease in selectivity at larger lengths which appears to be unnecessary for the alternately weighted experimental models.

I feel the simplicity of the model in terms of its selectivity has significant appeal in terms of providing confidence in the assessment something which is also reflected in the greater model stability. However, all presented models clearly struggle to fully explain the length compositional data with relatively small but strongly systematic cohort and length effects. I expect that this is due to the difficulty in appropriately modelling growth, or more precisely in estimating Lmax. In addition, despite these apparent difficulties for northern rock sole, the individual model estimates the cv for Lmax to be relatively low in contrast the cv between the models is roughly twice the within model cv.

Lack of contrast:

All model outputs show the same general stock development with little long-term contrast in the important metrics of F and recruitment. Fishing mortalities for the more compositionally weighted models appear to show no trend for either stock and convincingly demonstrates a high degree of synchrony in the inter-annual variation between the fisheries which seems appropriate given that it is essentially the same fishery. For northern rock sole inter annual variation appears unrealistically large for the two alternatively weighted models at maximum selection. However, the effective F, the proportion of fish that die, is very similar with much less interannual variation within all models even for northern rock sole. Taking into account the needed interannual variation in effort by a targeting fleet (Figure 9), does indeed suggest to me that the alternately weighted (17.2 a, b) models are less likely for northern rock sole and inconsistent with the (17.2 a, b) models for southern rock sole.

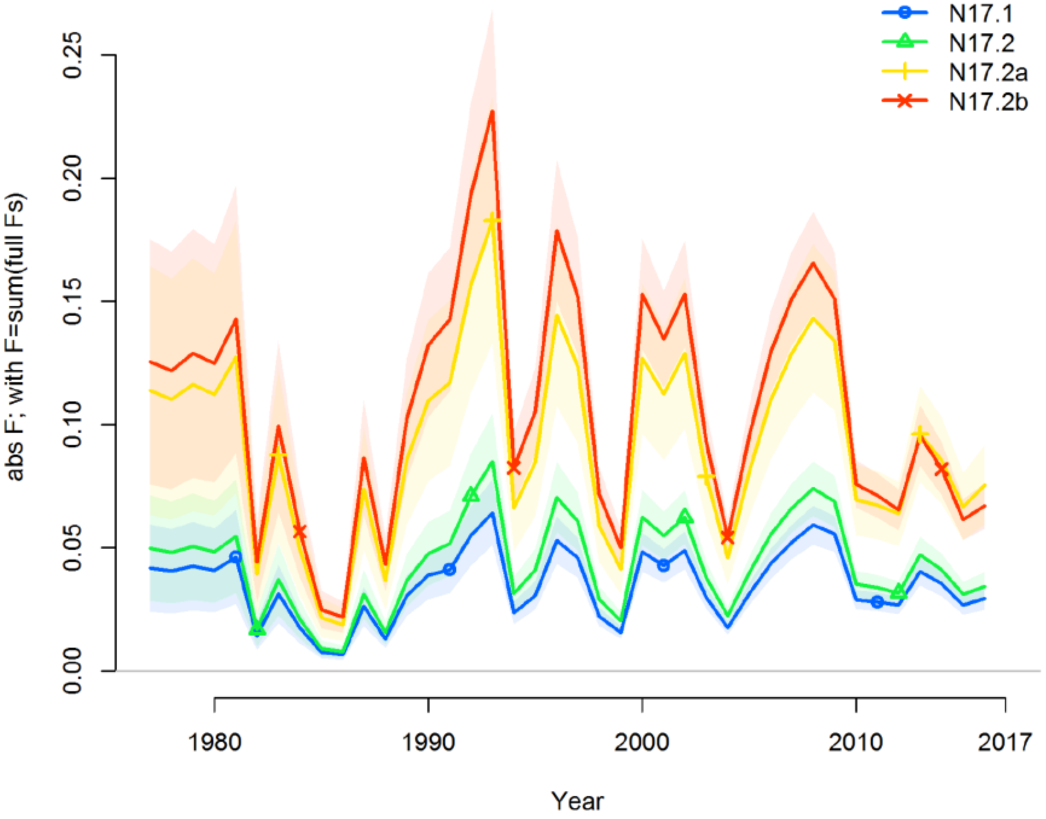


*Figure 9: Northern rock sole fishing mortality at full selection. With F=q*s*effort this would suggest unlikely interannual variation in effort for 17.2a and b models.*

A property of length-based catch-at-age models such as these is that when the faced with low signal to noise ratios they tend to converge to effectively biomass dynamic models. Scientifically, this is appropriate when there is little useful information in the compositional information. The issue from a management perspective is that they are still presented as catch-at-age models which allow for alternate, potentially less precautionary management strategies than for designated biomass dynamic models.

Critically with the transition from purely biomass based to catch-at-age is a smooth one with no established criteria for when alternate forms of management may be indicated. Characteristics of a predominantly biomass driven model are frequently the described lack of contrast, poor fit to compositional length data and difficulties in estimating selection parameter, and conflicting information in the various data sources. The lack of contrast/process signal means that often the models can be over-parameterised even if the parameter to data (number of samples) ratio is low. On the basis of the reviewed assessment report, I would identify the risk of over-confidence in the model output to be relatively high due to low parsimony, and assessment focus should be to investigate the modelling of selectivities/growth to improve the randomness of the length residuals, irrespective of their overall magnitude which is in the end largely determined by the choice of weighting used.

Model consistency:

Retrospective:

Northern rock sole has been suffering in some uncertainty in the estimation in the size of the last strong recruitment pulse which have just started to show up in the fishery. It seems highly likely that this discrepancy in the cohort strength is the result of the difficulty in reflecting the survey length compositional data, because the more numerous weak year classes are indicating increased recruitment estimates. Further contributing factors to the retrospective in SSB could be the impact of the year effect in the recent surveys applying extra leverage at the end of the timeseries. However, the scale of the retrospective bias is small and should not have a negative impact on management in this stock.

There is very little retrospective bias apparent in the Southern rock sole assessment, but the assessment is structurally identical and suffers from similar, though smaller, symptoms. The survey indicates the 2013 and 2014 cohorts to have been very strong so it is highly likely that this assessment will go through the same process of adjusting the recruitment strengths over the next five years; so a similar retrospective to that seen in the Northern stock in relation to the 2011 cohort is very likely.

Consistency of the models seems to have been predominantly determined on the basis of the analysis of the retrospective patterns. I agree both models are reasonably stable in their retrospective view of the management metrics. However, some of the parameters seem to indicate substantial re-estimation year to year. It seems much of the stability in the model output is caused by the low exploitation rate. With Z being close to M and M fixed it seems the model has few options other than altering recruitment to balance the dynamics retrospectively adjusting several parameter values to arrive at much the same result. To me this does not necessarily suggest a robust model under greater exploitation rates, and if these were to increase, model robustness would likely decrease significantly.

MCMC and jitter analysis:

The analyses represent a more comprehensive evaluation of robustness for the previously mentioned reasons. The preliminary results confirm that the model is not uniquely identified by the data. However, more thought needs to be put into how to most effectively set up some of these simulations so they are representative of real-world decision-making scenarios. Examples for this are setting appropriate thinning criteria for MCMC scenarios and deciding on the convergence criteria for jitter analysis. The results presented in this review tend to take a scattershot approach with a large proportion of the simulations representing very unlikely combinations of parameters which are largely uninformative on model robustness.

My conclusions are:

- The model output is largely robust and I would not expect to see substantial change in the near to medium future in SSB or F. Some changes to the estimates of recent recruitments are highly likely but unlikely to be deleterious to management given the broad age structure of the population.
- The models themselves do not reflect some population dynamic processes very well ,particularly growth and selectivity. This has knock-on effects on other parameters which the model is adjusting in a compensatory manner at present, but this is unlikely to remain the same under scenarios of heavier exploitation.

Critique of Review Process

I found this review process to be highly rewarding and interesting review and thank all the participants for their open and easy communication. The data collection presentations were very useful and universally appreciated by the panel. The reviewers' questions were given the necessary attention and opinions could be exchanged extensively and were received with an open mind. An excellent process.

Some issues that the CIE may wish to consider are:

- On-line meetings, while currently non-negotiable, do have some negative impacts on the review process. With all reviewers in Europe there was some limit to the time available. What I missed was the ability to go back to the assessor or data collection staff to confirm some details or have more general sidebar discussions which would normally take place informally during breaks, etc. The lack of such 'chats' also has an impact on the development of communication during a meeting as it is not possible to develop personal relationships and common thoughts. Time for the panel to do its formal work in the reduced time was not an issue.
- In this case a number of analysis and explorations could only be conducted at a very preliminary levels which means many of the 'conclusions' are really more 'expectations' since there was insufficient time to develop the necessary analyses. The situation was made worse by having a single assessor presenting all the assessments. Usually, it is possible for the assessors to do some work while one of the other assessments is being presented or discussed.

Appendix 1: Bibliography of materials provided for review

CIE Materials:

Link to google drive for CIE materials:

https://archive.fisheries.noaa.gov/afsc/refm/stocks/plan_team/2021_flatfish_cie/

List of documents provided:

1.) Draft agenda (LINK)
2.) CIE Statement of Work (LINK)
3.) Most recent stock assessments for BSAI Greenland turbot (LINK) and GOA northern and southern rock sole (LINK)
4.) Previous assessments for BSAI Greenland turbot (2018, 2016, 2015, 2014) and GOA northern and southern rock sole (2016, 2015, 2014, 2012, 2011, 2010)
5.) Link to all historic stock assessment and fishery evaluation reports (LINK)
6.) Stock assessment history for BSAI Greenland turbot (LINK) and GOA northern and southern rock soles (LINK)
7.) Groundfish fishery management plans for the Bering Sea and Aleutian Islands (LINK) and the Gulf of Alaska (LINK)
8.) Most recent North Pacific observer program sampling manual (2020)
9.) Recent paper on Greenland turbot archival tagging (LINK)
10.) Most recent ecosystem status report briefs for the Bering Sea (2020), Aleutian Islands (2020), and Gulf of Alaska (2020)
11.) Link to full ecosystem status reports (LINK), 2020 GOA Ecosystem Status Report (LINK)
12.) Link to full economic SAFE reports (LINK), 2020 Economic SAFE Report (LINK)
13.) Most recent stock synthesis user manual (LINK)

List of pre-recorded presentations (LINK to all presentations):

1.) Overview of the Observer Program and BSAI Greenland turbot observer fishery data
2.) Overview of the eastern Bering Sea bottom trawl shelf and slope survey (separate presentations) and BSAI Greenland turbot survey data
3.) Overview of the AFSC longline survey and BSAI Greenland turbot longline survey, tagging data, and recent manuscript on tagging data
4.) Overview of the GOA rock soles observer fishery data
5.) Overview of the GOA bottom trawl survey and GOA northern and southern survey data
6.) Overview of the AFSC aging methods and otolith data for BSAI Greenland turbot and GOA northern and southern rock soles

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

Gulf of Alaska Northern and Southern Rock Sole,
Bering Sea Aleutian Islands Greenland Turbot

January 25-29, 2021

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 stock assessments for Gulf of Alaska (GOA) northern and southern rock sole and Bering Sea and Aleutian Islands (BSAI) Greenland turbot provide the scientific basis for management advice considered and implemented by the North Pacific Fisheries Management Council (NPFMC). An independent review of these integrated stock assessments is requested by the Alaska Fisheries Science Center’s (AFSC) Resource Ecology and Fisheries Management Division (REFM).

The goal of this review will be to ensure that the stock assessments represent the best available science to date and that any deficiencies are identified and addressed. The specified format and contents of the individual peer review reports are found in **Annex 1**. The Terms of Reference (TORs) of the peer review are listed in **Annex 2**. Lastly, the tentative agenda of the panel review meeting is attached in **Annex 3**.

Requirements
NMFS requires three (3) reviewers to conduct an impartial and independent peer review in accordance with the PWS, OMB guidelines, and the TORs below. The reviewers shall have a working knowledge and recent experience in the application of stock assessment methods in general and with Stock Synthesis in particular. The chair, who is in addition to the three reviewers, will be identified and provided by the Alaska Fisheries Science Center (AFSC).

Tasks for Reviewers

1) Review the following background materials and reports prior to the review meeting:

Bering Sea and Aleutian Islands Greenland Turbot

Bryan, M.D., Barbeaux, S. J., J. Ianelli, D. Nichol, and J. Hoff. 2018. Assessment of the Greenland turbot (*Reinhardtius hippoglossoides* in the Bering Sea and Aleutian Islands. In Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region as projected for 2019. Section 5. North Pacific Fishery Management Council, Anchorage, AK.

Barbeaux, S. J., J. Ianelli, D. Nichol, and J. Hoff. 2016. Assessment of the Greenland turbot (*Reinhardtius hippoglossoides* in the Bering Sea and Aleutian Islands. In Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region as projected for 2017. Section 5. North Pacific Fishery Management Council, Anchorage, AK.
https://www.afsc.noaa.gov/REFM/Docs/2016/BSAIturbot.pdf

Barbeaux, S. J., J. Ianelli, D. Nichol, and J. Hoff. 2015. Assessment of the Greenland turbot (*Reinhardtius hippoglossoides* in the Bering Sea and Aleutian Islands. In Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region as projected for 2016. Section 5. North Pacific Fishery Management Council, Anchorage, AK.
https://www.afsc.noaa.gov/REFM/Docs/2015/BSAIturbot.pdf

Barbeaux, S. J., J. Ianelli, D. Nichol, and J. Hoff. 2014. Assessment of the Greenland turbot (*Reinhardtius hippoglossoides* in the Bering Sea and Aleutian Islands. In Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region as projected for 2015. Section 5. North Pacific Fishery Management Council, Anchorage, AK.
https://www.afsc.noaa.gov/REFM/Docs/2014/BSAIturbot.pdf

Gulf of Alaska rock soles

Bryan, M.D. 2017. Assessment of northern and southern rock sole (*Lepidopsetta polyxstra and bilineata*). In Stock assessment and fishery evaluation document for groundfish resources in the Gulf of Alaska as projected for 2018. Section 4. North Pacific Fishery Management Council, Anchorage, AK.
https://archive.fisheries.noaa.gov/afsc/REFM/Docs/2017/GOAnsrocksole.pdf

A’mar, T., Palsson, W. 2015. Assessment of northern and southern rock sole (*Lepidopsetta polyxstra and bilineata*) for 2016. In Stock assessment and fishery evaluation document for groundfish resources in the Gulf of Alaska as projected for 2016. Section 4. North Pacific Fishery Management Council, Anchorage, AK. https://archive.fisheries.noaa.gov/afsc/REFM/Docs/2015/GOAnsrocksole.pdf

A’mar, T., Palsson, W. 2014. Assessment of northern and southern rock sole (*Lepidopsetta polyxstra and bilineata*) for 2015. In Stock assessment and fishery evaluation document for groundfish resources in the Gulf of Alaska as projected for 2016. Section 4. North Pacific Fishery Management Council, Anchorage, AK. https://www.fisheries.noaa.gov/resource/data/2014-assessment-northern-and-southern-rock-sole-stocks-gulf-alaska

Additionally, two weeks before the peer review, the NMFS Project Contact will send by electronic mail or make available at an FTP site to the CIE reviewer any updated background information and reports for the peer review. In the case where the documents need to be mailed, the NMFS Project Contact will consult with the CIE on where to send documents. The CIE reviewer shall read all documents in preparation for the peer review.

2) Attend and participate in the panel review meeting. The meeting will consist of presentations by NOAA scientists, including the stock assessment authors, survey team members, and age and growth experts to facilitate the review, provide any additional information and answer questions from the reviewers.

3) After the review meeting, reviewers shall conduct an independent peer review report in accordance with the requirements specified in this PWS, OMB guidelines, and TORs, in adherence with the required formatting and content guidelines; reviewers are not required to reach a consensus.

4) Each reviewer should assist the Chair of the meeting with contributions to the summary report if required in the terms of reference.

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

Foreign National Security Clearance

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

Place of Performance

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

Period of Performance

The period of performance shall be from the time of award through April 2021. The CIE reviewers' 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	Deliverables and Milestones
Within two weeks of award	Contractor selects and confirms reviewers
Approximately 2 weeks later	Contractor provides the pre-review documents to the reviewers
January 25-29, 2021	Panel review meeting
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

All travel expenses shall be reimbursable in accordance with Federal Travel Regulations (http://www.gsa.gov/portal/content/104790). International travel is authorized for this contract. Travel is not to exceed $8,000.

Restricted or Limited Use of Data

The contractors may be required to sign and adhere to a non-disclosure agreement.

Project Contact(s):

Meaghan Bryan
Resource Ecology & Fisheries Management Division
NMFS| Alaska Fisheries Science Center
7600 Sand Point Way NE, Bldg. 4, Seattle, WA 98115-6349
Phone: 206-526-4694

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

Bering Sea and Aleutian Islands Greenland turbot

1. Evaluation of the ability of the stock assessment model for BSAI Greenland turbot, with the available data, to provide parameter estimates to assess the current status of Greenland turbot in the BSAI
2. Evaluation of the strengths and weaknesses in the stock assessment model for BSAI Greenland turbot
3. Recommendations for improvements to the assessment model.

Gulf of Alaska rock soles

1. Evaluation of the ability of the stock assessment model for GOA rock soles, with the available data, provide science advice to inform the management of rock soles in the Gulf of Alaska
2. Evaluation of the strengths and weaknesses in the stock assessment model for GOA rock soles
3. Recommendations for improvements to the assessment model.

Annex 3: Tentative Agenda

CIE Panel Review of Gulf of Alaska Northern and Southern Rock Sole, Bering Sea Aleutian Islands Greenland Turbot

TBD

Alaska Fisheries Science Center
7600 Sand Point Way NE
Seattle, WA 98117

Panel meeting January 25-29, 2021

Point of contact: Meaghan D. Bryan (meaghan.bryan@noaa.gov)

Appendix 3: Panel membership or other pertinent information from the panel review meeting.

Name	Program	Responsibility
Delsa Anderl	Age and Growth Program	Supervisor of otolith readers
Daniel Armellino	Fisheries Monitoring and Analysis	Review of rock soles in the observer program
Lyle Britt	Groundfish Assessment Program	Review of Bering sea shelf and slope bottom trawl survey and Greenland turbot data
John Brogan	Age and Growth Program	Review of aging for Greenland turbot and rocksoles
Katy Echave	Marine Ecology and Stock Assessment Program	Longline survey tagging data
Jim Ianelli	Status of Stocks and Multispecies Assessment	Historical stock assessment
Sandra Lowe	Status of Stocks and Multispecies Assessment	Supervisor of stock assessment authors
Pat Malecha	Marine Ecology and Stock Assessment Program	Supervisor of longline survey and tagging
Wayne Palsson	Groundfish Assessment Program	Review of Gulf of Alaska bottom trawl survey and rock soles data, program supervisor
Raul Ramirez	Fisheries Monitoring and Analysis	Review of Greenland turbot in the observer program
Kevin Siwicke	Marine Ecology and Stock Assessment Program	Review of longline survey and tagging for Greenland turbot

References:

Bryan, M.D., S. J.Barbeaux, J. Ianelli, D. Nichol, and J. Hoff. 2018. Assessment of the Greenland turbot (Reinhardtius hippoglossoides in the Bering Sea and Aleutian Islands. In Stock assessment and fishery evaluation document for groundfish resources in the Bering Sea/Aleutian Islands region as projected for 2018. Section 5. North Pacific Fishery Management Council, Anchorage, AK.