Independent peer review report for the assessment of Bering Sea Aleutian Islands Greenland Turbot, Gulf of Alaska Northern and Southern Rock Sole

Dr. Anders Nielsen

Completed for the Center for Independent Experts (CIE)

June 2021

Executive Summary

The Bering Sea Aleutian Islands (BSAI) Greenland Turbot, Gulf of Alaska (GOA) Northern and Southern Rock Sole are being assessed with three separate assessment models. The three models being used are all configured in the stock synthesis general assessment model framework and the three models are somewhat similarly configured and have somewhat similar strengths and weaknesses.

The available data are well documented of good quality and the assumptions made in the assessment models are overall reasonable. The models are not describing all details of the data, but they are describing the main trends.

All three models are all configured to be very flexible with respect to the selection curves assumed. It is obviously desirable to allow this flexibility within the models, but there are strong indications from the model diagnostics that with the available data the models are struggling to converge to unique solutions. A first priority should be to investigate this convergence fully, which will likely lead to a (possibly small) reduction in the flexibility of the selection patterns (especially for BSAI Greenland Turbot).

The details of the configurations include fixed parameters, prior distributions (deviances with fixed variances), assumed variances or effective sample sizes. These things are common in assessment models, but they do obstruct the models' ability to correctly quantify the uncertainties of the estimated time series.

Simplifying the selection patterns could possibly lead to models where it would be possible to estimate the catchabilities relating the absolute biomass indices to the abundances. For BSAI Greenland Turbot the practice of using estimates from an older model run is not recommended, because it uses the first part of the data twice. For the two rock sole stocks the assumption of equal density in trawlable and untrawlable areas would be less critical if the catchability did not have to be fixed.

The assumed 50/50 split of the catches between northern and southern rock soles could also be reconsidered. Assuming the same catch known without uncertainty could lead to an unwanted and unrealistic correlation between the two stocks' development and to an underestimation of the uncertainties. An interesting long-term solution could be a joint model where the split-fraction was a part of the model. A possible short-term improvement could be to assume some level of observation noise in catches and possibly use the split data, which is available since 1997.

The review meeting was efficient and well organized by Kalei Shotwell and Meaghan Bryan. However, having an assessment review online is not a good substitute for an actual review meeting.

Background

The review workshop of the Bering Sea Aleutian Islands Greenland Turbot, Gulf of Alaska Northern and Southern Rock Sole assessments was requested by the Alaska Fisheries Science Center's (AFSC) Resource Ecology and Fisheries Management Division (REFM). The meeting was conducted via a five-day online meeting (5-9 April 2021) and a small pre-meeting a few days prior (where the technical setup was tested and agenda briefly discussed). Prior to the meeting the written material and seven presentations covering the data collection (visuals and audio) were made available to the reviewers. At the online meetings the reviewers were given the opportunity to ask questions to the pre-provided presentations. The lead assessment scientist Meaghan Bryan presented all details about the three assessments and answered all questions from the reviewers and provided additional explorative assessment runs and additional model diagnostics. The relevant documents (see Appendix 1) were made available in ample time prior to the meeting. The meeting was carefully prepared and well organized by Kalei Shotwell and Meaghan Bryan who made the proceedings run very efficiently. The goal of such a review meeting is to strengthen confidence that the assessments are scientifically sound and that the results are reliable. This report documents the independent review of Center for Independent Experts (CIE) reviewer Anders Nielsen (see appendix 2 for statement of work).

Description of the individual reviewers' role

This reviewer has independently read the assessment reports, and all supplementary documents deemed necessary in preparation for this review, participated in an online pre-meeting, participated actively in online review meetings (5, 6, 7, 8, and 9 April 2021) identified key issues in the assessment and validation, suggested guidance, and independently authored this review report.

Findings for each TOR:

To ensure that all terms of reference are covered and that comments are interpreted with reference to the correct terms, the terms are listed (with gray highlighting) with corresponding reviewer comments following. Since the models used for BSAI Greenland turbot and GOA rock soles are fairly similar configurations of the same modelling framework stock synthesis (SS), there will be some repetition of comments from one stock to the other. I decided that this was preferable in order to have complete report sections for each stock.

However, first there is a common comment relating both TOR 1's for all three stocks. The models presented for both Gulf of Alaska rock sole stocks, and for Bering Sea and Aleutian Islands Greenland turbot are all based on the very general and highly configurable stock assessment model `Stock Synthesis'. Stock synthesis is among the most applied stock assessment models in the US and in the world. It is part of the NOAA Fish and Fisheries Toolbox (Fish-Tools https://nmfs-fish-tools.github.io/). Stock synthesis has been validated in numerous peer reviewed assessments (e.g. SEDAR 54: HMS Sandbar Shark, SEDAR 39: Atlantic Smooth Dogfish, and SEDAR 44: Atlantic Red Drum), in peer reviewed scientific journal papers (e.g. Method & Wetzel 2013, Punt & Maunder 2013, and Zhu et al. 2016), and in meetings dedicated to evaluate assessment models (e.g. World Conference on Stock Assessment Methods for Sustainable Fisheries, 2013, Boston; Workshop on Recent Advances in Stock Assessment Models Worldwide, 2010, Nantes; and many Center for the Advancement of Population Assessment Methodology (CAPAM http://www.capamresearch.org/) workshops). Recently the source code for the stock synthesis program has been made publicly available at github (https://github.com/nmfs-stock-synthesis). All of these things strengthen the confidence that the code is correct and has been thoroughly validated to be robust in a high number of cases.

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

The stock assessment model for BSAI Greenland turbot is configured in the flexible assessment model framework stock synthesis (see general comment above). The model does have the ability, with the available data, to provide parameter estimates and to assess the status. However, every part of stock synthesis can be configured in many different ways. Stock synthesis has been used to assess BSAI Greenland turbot for a long time (since 1993), so there is a lot of experience to configure it for this stock.

The data used to assess the BSAI Greenland turbot stock are catches split into longline and trawl. In the earliest part (1960-1970) of the time series the catches included a different stock (arrowtooth flounder), so for the first few years (1960-1964) the ratio from the following period (1965-1969) of the Russian fleet was used to split. This approach seems reasonable and isolated to the first part. Length samples for both catch types are available from 1980 to 2020 with very few years missing. Three surveys are used in the assessment. The Eastern Bering Sea (EBS) slope and shelf surveys and an Alaska Fisheries Science Center (AFSC) longline survey. The shelf survey provides an index of the juvenile part of the populations and the slope survey provides an index of the older juvenile and adult part of the population. The shelf survey is available yearly from 1987 including length compositions, but due to survey consistency issues, the slope survey index is only judged usable in certain recent years (2002, 2004, 2008, 2010, 2012, and 2016), whereas the corresponding length composition are used in those years and in six additional years (1979, 1981, 1982, 1985, 1988, and 1991). The age information from the shelf survey was included in the form of mean-length-at-age (this approach does to some to some degree re-use the length information already included in the model, but this is likely not important). The longline survey available from 1996-2016 is combined from two surveys taken in alternate years.

The population model is configured to be sex-specific. The natural mortality (M) is assumed fixed and common for both males and females. It is convincingly illustrated that the assumed M is within the range of a differently derived alternatives. The length weight relationship is assumed fixed, but sex-specific. The Von-Bertalanffy growth parameters are sex-specific and estimated within the assessment model. The distribution around the growth curve is determined by assigning a fixed length CV's at two ages (1 and 21) and interpolating the remaining CV's. The maturity-at-length function is externally estimated and kept fixed in the assessment model. A Beverton-Holt stock-recruitment function is assumed where the steepness is fixed (h=0.79) and the initial recruitment estimated. The actual recruitment estimates are allowed to deviate from the Beverton-Holt curve with autocorrelated log-scale deviations with fixed standard deviation but estimated (prior penalized) correlation parameter. The population part of the model appears well balanced to describe the population and at the same time be sufficiently constrained to be estimable with the available data. Where the quantities are not estimated within the model the choices are well reasoned and derived from cited sources. Of these the choices which appear most arbitrary are the assumed standard deviations (e.g., on growth envelope and on recruitment deviations).

The observation model is size- and sex-specific. The catches from the two catch-fleets are assumed to be known without observation noise. The total indices from the surveys are assumed log-normally distributed with standard deviations carried over from the survey index calculation. The composition data from both the two catch fleets and the shelf and slope surveys are assumed to follow multinomial distributions (with assigned effective sample sizes). The selectivity assumed to match the composition observations are very flexible. Sex-specific, time-blocked double-normal selection curves for the two catch fleets and for shelf survey. A sex-specific, time-blocked logistic selectivity the slope survey and fixed logistic for the longline. The overall catchability is estimated for the longline survey but fixed for the slope survey (q=0.574) and shelf survey (q=0.616). The two fixed values are based on a previous model run (in 2015), so in fact re-using parts of the data also used in this model. The observation model is set up to be very flexible and detailed with respect to the selection pattern. Fixing the standard deviations to the standard deviations from the survey calculations is reasonable to capture the year-to-year variation in uncertainty. Fixing the remaining variance parameters (including the fixed sample sizes) is standard practice within stock synthesis with the aim to assign the appropriate relative weight somewhat subjectively to the different data sources. If the relative weighting is assigned correctly it leads to correct estimates of parameters and stock status. It does however not lead to objective variance estimates of those relevant quantities.

2. Evaluation of the strengths and weaknesses in the stock assessment model for BSAI Greenland turbot

The main strengths of the model are that it is able to accommodate all the different sources of data, the code is extensively applied and thereby tested, and that BSAI Greenland turbot has been assessed by different configurations of stock synthesis models since 1993, so there is a lot of experience. The main weaknesses are that it appears that some parameters are not well estimated and some of the fixed quantities.

The way the catchabilities are fixed by using the estimates from a previous (2015) version of the model for BSAI Greenland turbot with the old data series as fixed input to this mode is unusual. The estimates do not appear to be unrealistic. If it was just a “time saving” or “model stabilizing” trick, then it would not be much of an issue, but at the review meeting it was attempted to estimate it with the current data, which resulted in unrealistic estimates. The previous model was possibly more restricted in other parts, which allowed the catchabilities to be estimated. The double use of the first part of the data series should be avoided. This issue is possibly linked with the possibly too flexible selectivity issue.

Some parameters are fixed (effective sample sizes, and variance parameters (including those assumed to be zero)) and such fixed variance parameters will directly influence the estimated uncertainties of all estimated quantities. The focus is largely on assigning the appropriate relative weights to the different information sources, which is important to get the correct model estimates. However, the absolute weights are important for all estimated uncertainties, and assigning fixed parameters (e.g., catchabilities and fixed variances) does hinder the model in correctly propagating the uncertainties from the measurement noise in the observations onto the uncertainties on estimated model parameters and quantities of interest (such as stock status). Using the standard deviations from the survey calculation is a good step in using the uncertainty in the data (rather than just assigning relative weights).

At an overall level, the model is able to describe all the data sources (fig. 5.14, 5.15a, and 5.16), but the size compositions of the trawl catch fleet and the slope survey does not match the peak of the male compositions (5.16). Also looking at e.g., the composition residuals (e.g., fig. 5.17) it appears that some cohort effects are not well described by the model. It also appears (top frame fig. 5.17) that the selectivity is changing gradually (e.g., from 1984 to 1990) and hence not necessarily well approximated by a few breakpoints.

The estimated selectivity patterns appear to have very dramatic shifts from one period to the next (e.g., fig 5.18). There are some actual reasons for this (different fleets). Some of the estimated selection patterns appear erratic or even implausible (e.g., right column of figure 5.19).

A very flexible selection pattern is desirable, but after seeing the results, it seems that the assumed functional forms for the selectivity for BSAI Greenland turbot may be too flexible compared to what can actually be estimated from the observations.

There is a retrospective pattern in the most recent years for SSB, but less so for fishing mortality (5.27); both are acceptable. The retrospective bias for recruitment is large, so the last year's estimate of recruitment should be used with caution. The retrospective analysis does indicate that the overall estimates of stock status (historic and current) is reliably estimated. This indicates that the model is able to provide stable estimates of the quantities of interest.

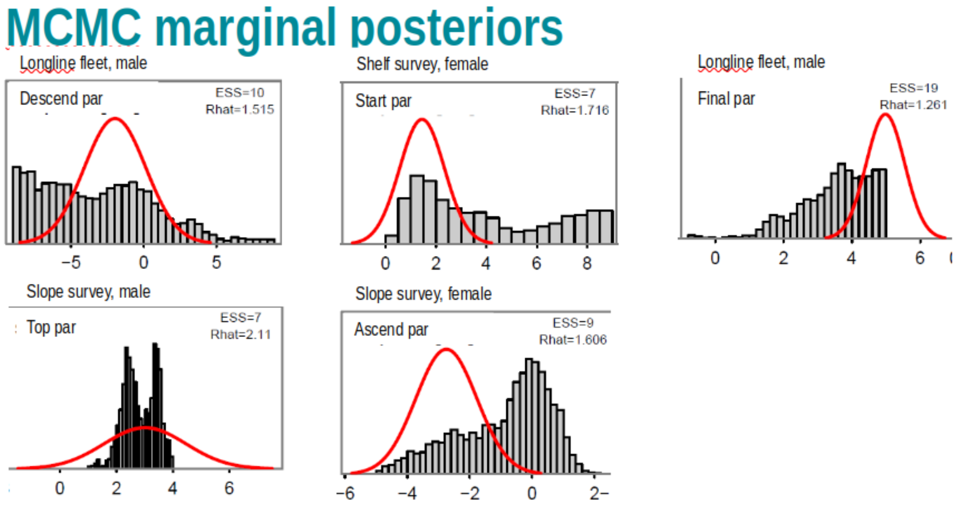

Figure 1: Some MCMC diagnostics from the review meeting. Red curve is Hessian based asymptotic normal distribution. Bar chart is MCMC simulated distribution for some selected selectivity parameters.

At the review meeting the robustness was further investigated by MCMC methods. MCMC can also illustrate if the normal distribution based asymptotic results that are standard practice to summarize uncertainties on estimated quantities are not able to capture the uncertain distributions for some estimates. For most parameters it appeared that the MCMC distribution was fairly identical to the approximate normal distribution, but for a few parameters the MCMC distribution was deviating. It can be a bit difficult to draw a conclusion from this because it can be: a) because the MCMC procedure did not reach convergence b) The normal distribution is a poor approximation for a given parameter, c) Some model parameters are unidentifiable. Furthermore, regularization priors or bounds were used to prevent the MCMC algorithm from suggesting values that would crash the algorithm. From looking at the graphs it was however quite clear that the problem was either a) or c) because the bar charts in figure 1 does not look like plausible error distributions for the model parameters (when it happens that the asymptotic normal distribution is a poor approximation an asymmetric distribution may be expected, but not one where the highest value is at the boundary, or one with multiple peaks).

A more direct way to evaluate if the model convergence is stable, and a unique global solution is obtained, is the so-called jitter analysis. Here the model is initialized by a range of different initial values, and it is verified that the model converges to the same value. I requested the jitter analysis.

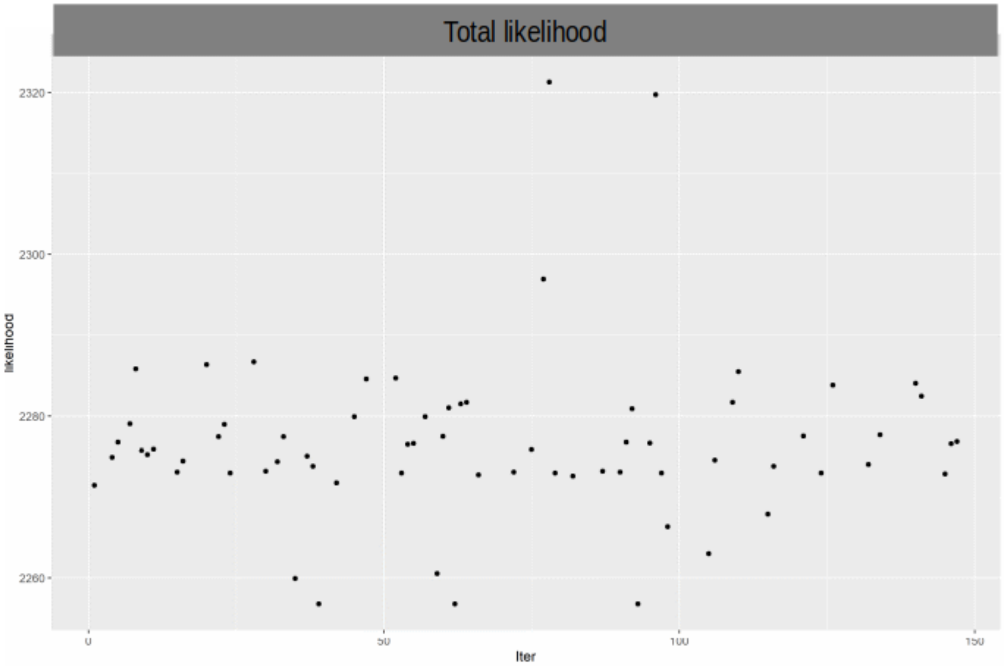

Figure 2: Jitter analysis for the BSAI Greenland Turbot model. Negative log likelihood of 68 converged runs (out of 150 jitters). The negative log likelihood for the base run was 2275.83.

The jitter analysis showed that the model had difficulty converging 68 out of 150 converged, which is not a big problem as long as it is clear from the model diagnostics output (e.g., gradient) when it has converged and when it has not. If an assessment model does not converge in a real applied case, then a different set of initial values can be applied to make it converge.

The jitter analysis further showed that the convergence was not to a unique global minimum even when only considering the runs which are diagnosed (by the model) as converged (figure 2). This is a more problematic issue, because for any real model run, we have no way of knowing if the model is stuck in a local optimum, and hence if we would draw different conclusions if we had initialized the model slightly differently. Considering the negative log likelihood value for the base run (2275.83) it appears that the base run did not reach the optimum (minimum of the negative log likelihood), because at least eight runs resulted in a better likelihood value (figure 2). Also, it should be noted that the likelihood values differ by non-trivial values.

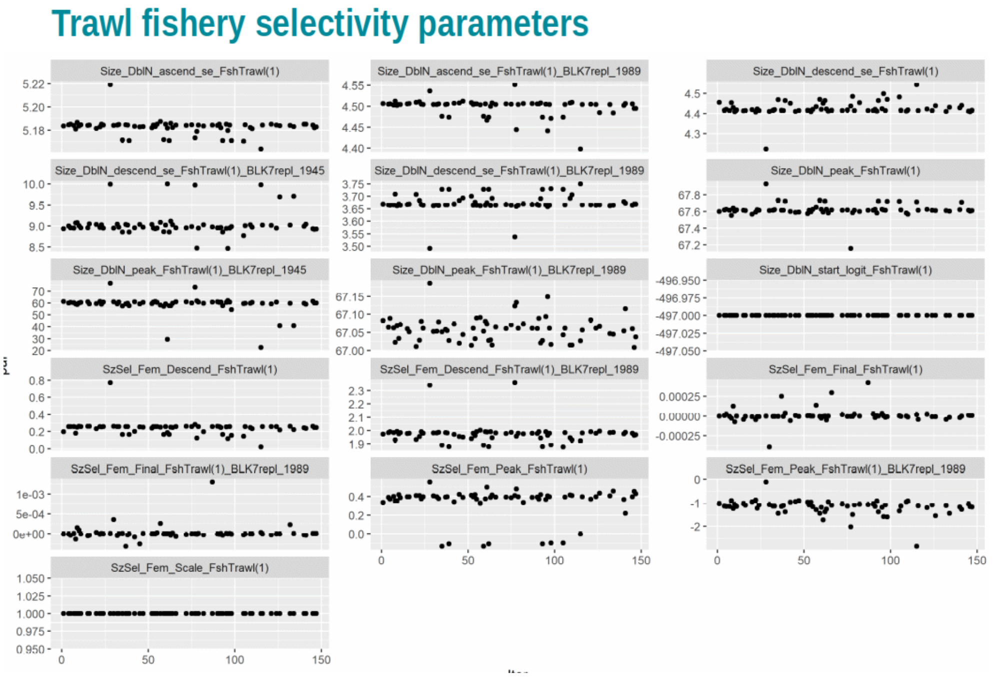

Figure 3: Converged jitter estimates of some selection parameters.

The jitter/MCMC analysis show that some parameters (see trawl selection parameters in figure 3) are not uniquely identifiable, or at least that the model declares convergence before the global minimum is obtained. This is the pattern that is typically seen if one or more parameters are able to compensate for each other. However, if these parameters are unimportant for the estimated model parameters of interest (e.g., stock status), then it would not be too problematic. This could for instance happen if it only affected a parameter that was used to accommodate a small part of the data set, which has almost no influence.

From the estimated spawning stock biomass in each of the converged jitter runs (figure 4) it is seen that the different converged runs do provide different estimates of the spawning stock biomass. In absolute terms the estimated SSB time series are quite different. In relative terms they are not so different, so all converged jitter runs show the same relative development of the stock over time.

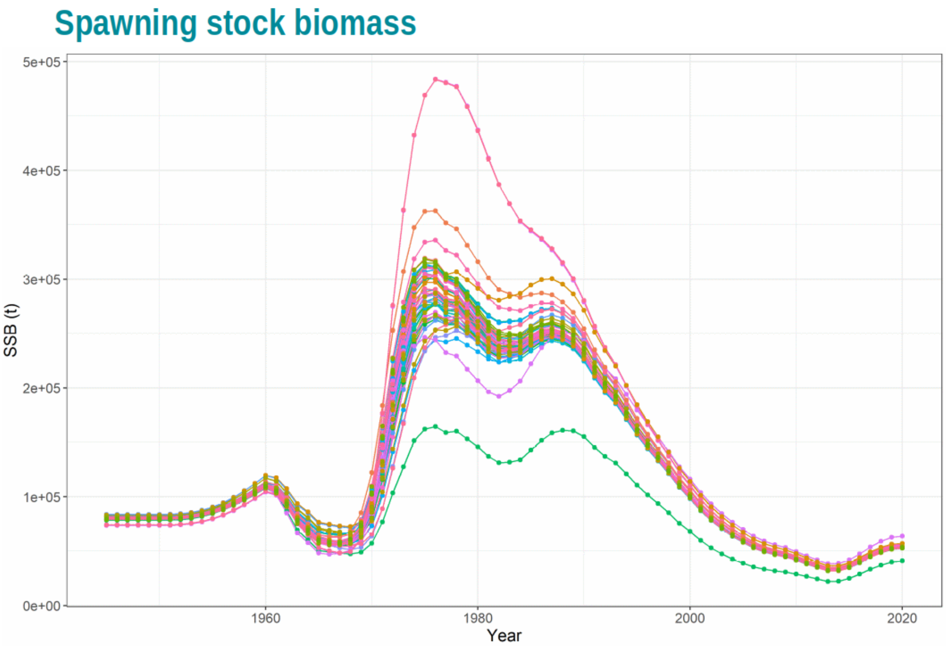

Figure 4: The spawning stock biomass as estimated from the 68 converged jitter runs.

3. Recommendations for improvements to the assessment model.

The first priority should be to stabilize the convergence of the assessment model. This will likely involve simplifying the model. The part of the model that is most flexible is the different sex- fleet- and time-specific selection functions. The double normal selection curve and the offset mechanism appear to be difficult to estimate from the available data, so a possible way to obtain a stable convergence for BSAI Greenland turbot data could be to start from a model with very simple selection curves everywhere e.g., logistic selectivities (verify stable convergence). This simple version will likely not describe the observations well. Once the model is simplified enough to achieve stable convergence it can be built back one step at a time to better describe the observations. If stable convergence is checked with every small enhancement of the model, then part that is causing the problem can be identified and possibly be solved by fixating (or otherwise restricting) hopefully very few model parameters. A few other things to consider are.

The procedure of using an old model run with the first part of the same data series to provide estimates of an important model parameters, such as the survey catchability could be reconsidered. Hopefully, when the convergence has been stabilized, it will be possible to estimate the catchabilities reliably within the model.

There are a few challenging issues in the first part of the timeseries (assumed split of the catch, fleets leaving the fishery, and less data available), so consider making an explorative assessment, where this period (e.g., first 20 years) was left out, simply to verify that the conclusions are not solely based on the more uncertain period.

Whenever the fishery has a doming selection, it can be a good idea to illustrate the “Cryptic biomass” as estimated by the assessment model. Just to see if it all of a sudden start to increase without any plausible reason.

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.

Stock Synthesis is one of the most general and complex assessment models, which is an advantage because it is applicable in many different scenarios and is able to accommodate many different types of observations. The many possible ways to setup and configure stock synthesis also increases the difficulty and knowledge required to operate the model correctly. The detailed discussions of the different options within stock synthesis at the review meeting clearly demonstrated that the models for the two rock sole stocks were configured with full knowledge of the options within stock synthesis.

The northern and southern stocks are similar in the data types. The data used in each of the models are: Total catches back to 1977, length compositions of catches back to 1997, survey in 11 of the 22 years from 1997 to 2017 with corresponding length compositions and conditional age-at-length. Age compositions are available, but not used directly.

The data available for the two rock sole stocks are overall sufficient to inform the management about the state of the stocks and provide scientific advice. This is partly because the stocks are not strongly targeted. The stocks are a major part of the so-called "shallow-water flatfish" stock complex, and in the history of managing this stock complex the total allowable catch (TAC) has never been caught (rarely has half of the TAC been caught). There are however a few problematic parts to the available data.

The total catch of rock sole is split by a fixed ratio in all years. 50% is assigned to the northern rock sole stock and 50% is assigned to the southern rock sole stock. The purpose of this splitting is to conduct independent assessments of the two stocks, which now have identical total catch time series, which are further assumed to be without observation noise for each stock. The observer program has identified the two species since 1997, so from 1997 these data could potentially be used instead of the fixed 50%. They appear to be centered around 50%, but with large variations from year to year (figure 5).

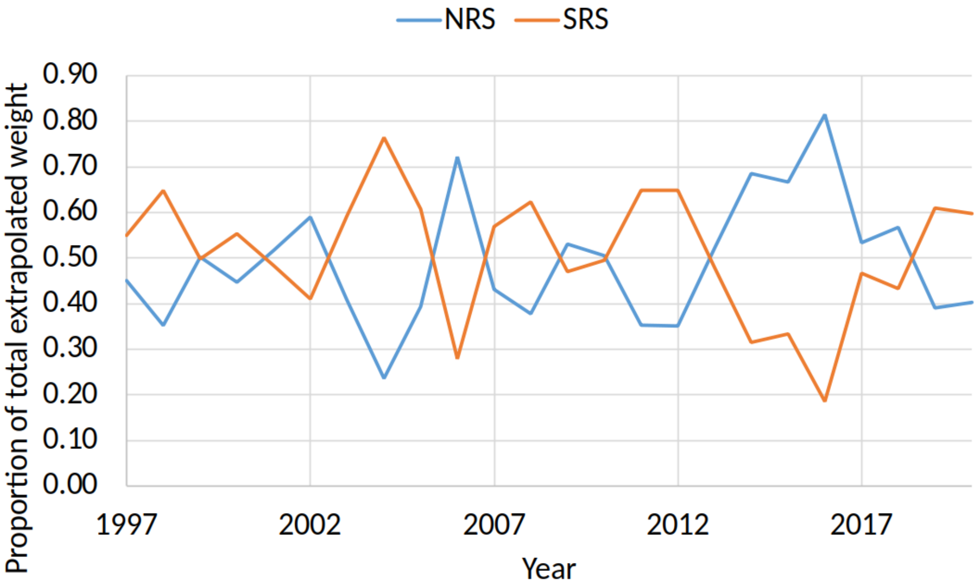

Figure 5: Observed split between northern and southern rock sole (graph from assessment presentation at review meeting).

The observed growth pattern, which is observed by observing age-at-length from the survey, shows an interesting pattern (e.g., figure 4.7). It appears that there is a relatively narrow relationship between age and length for lengths less than ca 35cm, and then a shift for lengths larger than ca 35cm where the ages are relatively more variable. From the age-reader statistics, and statements made at the meeting, it appears that age reading is very reliable for these two stocks, so the issue does not appear to be related to age reading uncertainty. Some flatfish are known to skip spawning in some years, and as a consequence be able to direct more energy to growth, and such behavior could possibly explain the issue.

The survey is used to calculate an absolute abundance index by scaling up the swept area to the total area. The survey only monitors the trawlable area, and the swept area method assumes that the species is evenly distributed in the trawlable and untrawlable areas. This is a strong assumption, because the same things which make an area untrawlable (e.g., rocks or strong currents) could make it attractive or unattractive as habitat for rock sole. If that is the case, then the absolute abundance index will become biased. Looking at the distribution maps presented at the review meeting it could appear that rock sole are more abundant in areas near untrawlable areas (see e.g., the sub-area near Kodiak Island (figure 6)), but this is difficult to judge from such images. The relative abundance in untrawlable areas is an active research area and could add very valuable information to these two assessments.

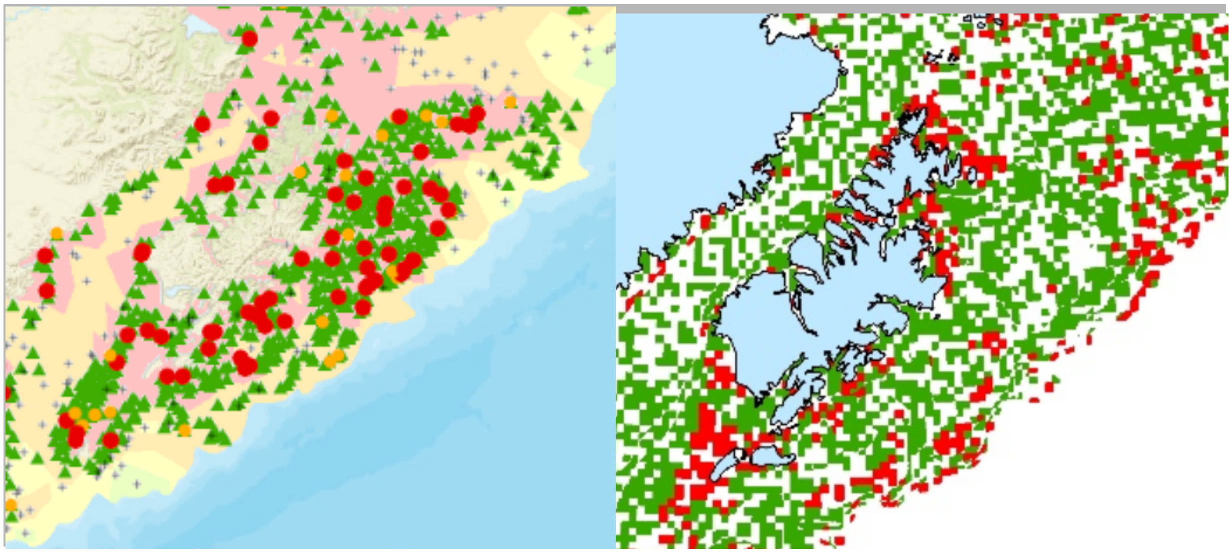

Figure 6: Part of distribution area (around Kodiak Island). Left frame: Southern rock sole distribution, green triangles indicate species extent all years red dots high intensity in 2019. Right frame: Green squares indicate trawlable stations and red squares in indicate untrawlable. (Graph partially cut from data presentation at review meeting).

2. Evaluation of the strengths and weaknesses in the stock assessment model for GOA rock soles

The assessment model (stock synthesis) is able to accommodate the data types available for the assessment of the two rock sole stocks. The two assessment models are configured very similarly (and not too dissimilar to the model for BSAI Greenland turbot described above):

The models are sex- and size- specific. The natural mortality of females is assumed fixed at 0.2, but the male natural mortality is estimated within the assessment model. New to me, and definitely worth highlighting, is the presented comparison to a lot of alternatively derived M-estimates, which are collected in an online app (http://barefootecologist.com.au/shiny_m). This is a very good idea and clearly illustrated that both the assumed and estimated values are within the range of plausible values for stocks with similar biological parameters. Sex-specific von Bertalanffy growth curves were estimated including size-distribution envelope and estimates appear reasonable. The stock-recruitment relationship used is a Beverton-Holt type, but with steepness fixed at 1, which essentially gives a constant curve. Penalized deviations around the stock-recruitment curve were assigned a log-scale fixed standard deviation of 0.6 (and allowed an initial offset value). Maturity-at-age (stock specific) and weight-at-lengths (common) are included as fixed inputs.

The observation models assume that the total catch for each stock is known without observation noise. The surveys are scaled (by inverse swept-area fraction) into an absolute index of abundance, and the variance estimates of this index calculation are used in the assessment models. The overall catchability relating the abundance indices to the true abundances for the two stocks are assumed to be one. The Length compositions of the catches and surveys are assumed to follow multinomial distributions. The selectivity curves are assumed to be asymptotic for the surveys and domed for the fishing fleet. Both the survey and fishing selectivities are sex-specific. The selectivities are assumed to be the same in all years. Few details are provided (distributional assumptions and such) about using conditional age-at-length, but the paper Lee et al. 2019 was helpful.

At the models are describing the observations at an overall level (figure: 4.11 (red), 4.12 (top right), and 4.18 (b,d,f,h) for northern rock sole and 4.26 (red), 4.27 (top right), and 4.33 (b,d,f,h) for southern rock sole), but even at the overall level there are some patterns of mismatch. The mismatch in overall match in the length compositions is different for the two stocks. For northern rock sole the model generally predicts too few in the center of the length distribution for males in the survey, where it for southern rock sole model does not have the same issue, but a less pronounced similar issue for the females. A more detailed year by year inspection of the residuals show similar patterns to a smaller or larger degree in almost all years, and also shows that the conditional age-at-length observations are underpredicted in some years.

The model for northern rock sole has a very consistent retrospective bias for the spawning stock biomass estimates (figure 4.25 a), with a Mohn’s rho of about 14%, which is high, but the stock is at a high and stable level, so not an immediate concern. The model for southern rock sole does not have a consistent retrospective bias for spawning stock biomass estimates. Neither of the models has consistent retrospective bias for fishing mortality or recruitment.

Different options were explored for the composition data (with or without using groundfish survey length compositions and with different adjustment for effective sample sizes). Based on retrospective performance and on critical evaluation of the realism of the adjusted effective sample sizes it was concluded that the unadjusted model including length compositions from the groundfish survey was preferable. These different model options also served as sensitivity runs and nicely illustrated that the main important estimates and conclusions from assessment are robust to some adjustments of the model formulation (4.22 and 4.37).

It is a bit unusual to have two independent assessments which have identical catches (a consequence of the 50/50 split). It is further seen that the observed fraction varies over time (to an 80/20 split when it is most extreme), which is likely partially due to observation noise and partly due to the true varying over time. Therefore, it seems unrealistic to assume that both of the catches are observed without observation noise.

The robustness of the model and the appropriateness of the asymptotic normal distributed estimation noise distributions was investigated at the review meeting via MCMC simulations. For both stocks the MCMC diagnostics appeared to be converged and for all, except a single parameter for each stock, the MCMC densities and the asymptotic normal densities were very similar. The single parameter that deviated was a poorly determined logit-scale parameter related to the female fishery selection pattern. This should be looked into but does not appear to be influential.

To further investigate if the model convergence was stable a jitter analysis was requested at the review meeting. In a jitter analysis the model is initialized by a range of different initial values, and it is verified that the model converges to the same value. The jitter analysis showed that the models are not easy to make converge for the northern rock sole 19 out of 100 runs did not converge and for the southern rock sole 45 out of 100 did not converge. The low convergence rate is annoying in practice, but not a major concern as long as it is clear from the model convergence criteria if a given model run converged or not.

For the northern rock sole model all the runs diagnosed as converged had final negative log likelihood values within 0.01 from each other, which means that in terms of fit to data the solutions are identical, but when looking at the corresponding model parameters of the converged runs we see that they are not the same for all runs (figure 7). This would indicate that some model parameters are able to compensate for each other such that the likelihood would become exactly the same. This is surprising, because the MCMC analysis did not indicate that. This jitter analysis was done during the meeting and under considerable time constraints, so possibly this is a false warning.

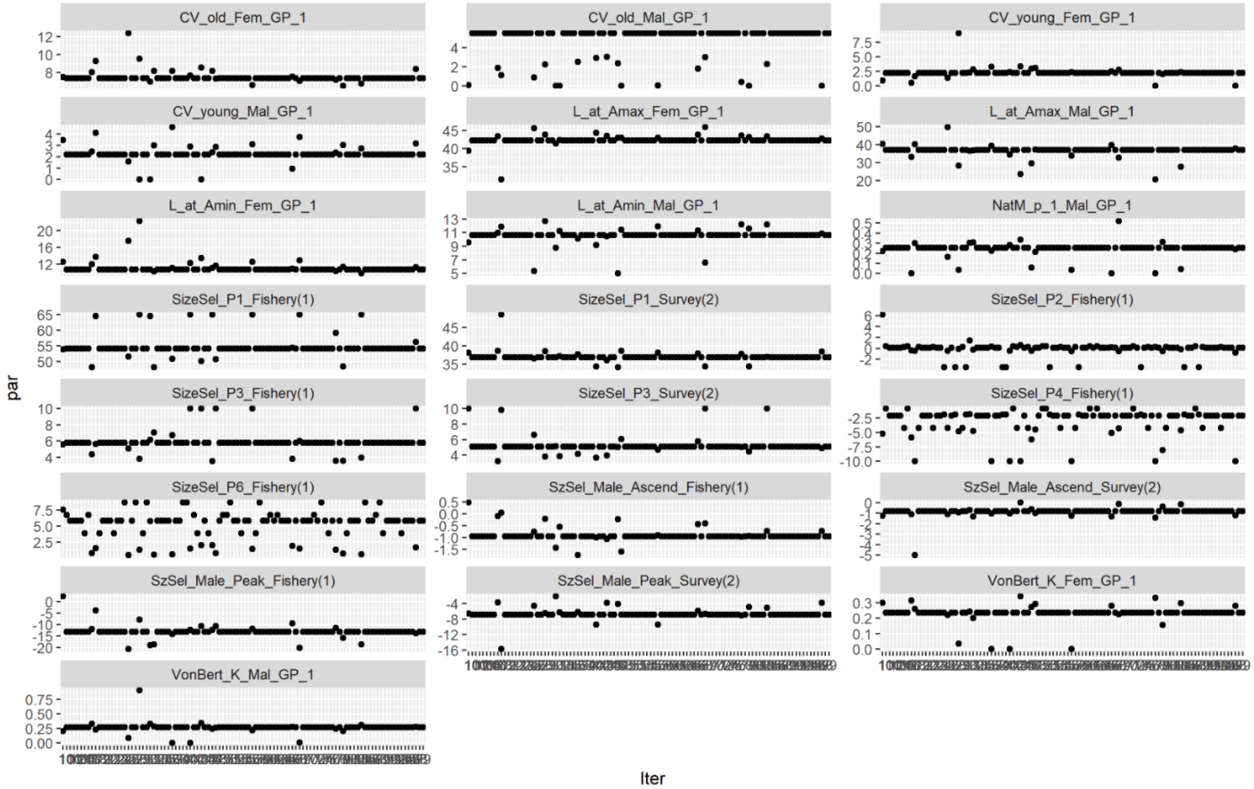

Figure 7: Parameters for northern rock sole converged jitter model runs.

For the southern rock sole model all the runs diagnosed as converged, except one gave the same final negative log likelihood, and the model parameters were also identical, except from a few outliers.

3. Recommendations for improvements to the assessment model.

The jitter analysis should be revisited, and it should be verified that for both stocks the model returns estimates of the quantities of interest (spawning stock biomass, fishing mortality, and recruitment) that are the same for all model runs diagnosed as converged (from a number of different starting values). Based on the retrospective runs, the sensitivity runs, and the MCMC diagnostics this is not expected to be a problem (but it should be verified). If it turns out to be a problem, it can likely be isolated to a few parameters, and it should hopefully not be too difficult to identify and solve (restrict or eliminate affected model parameters).

Since the fit to the conditional age-at-length is problematic (as seen in figures 4.18 and 4.33) and there could be more general problems with the use of this data type that extends outside this assessment (Lee et al. 2019) it could be considered to use the age compositions data directly in the assessment model.

The assumed 50/50 split giving the exact same catch series in both assessments, which is further assumed to be without observation noise, does sound problematic. At the very least it will underestimate the added uncertainty originating from split. A first step could be to allow some uncertainties in the catch time series to capture this uncertainty. A long-term goal could be to develop a joint model for the two stocks where the split fraction (process) was estimated internally. Such a model would capture the correlated coupling in the two catches (that they have to sum to the total catch).

The untrawlable areas seem to be an important issue for these assessments. Currently it is assumed that rock sole is equally abundant in trawlable and untrawlable areas in the survey calculation, and again when fixing the catchability parameter to 1 in the model. This is a strong and as far as I have heard, it is an unsubstantiated assumption. Relative abundance in the untrawlable is currently being researched and this research should be encouraged. Further it could be investigated if it is possible to estimate the survey catchability within the model (possibly at the cost of simplifying the models elsewhere).

Comments on the review process:

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

References:

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

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

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

H Lee, KR Piner, IG Taylor, T Kitakado, 2019. On the use of conditional age at length data as a likelihood component in integrated population dynamics models. Fisheries Research.

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)

12.) 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

April 5 – 9, 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) Prior to the peer review, the CIE reviewers will participate in a test to confirm that they have the necessary technical (hardware, software, etc.) capabilities to participate in the virtual panel in advance of the review meeting. The AFSC NMFS Project Contact will provide the information for the arrangements for this test.

3) 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.

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

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

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

Place of Performance

This review will be conducted via virtual meeting software.

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
April 5-9, 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

No travel is necessary, as this meeting is being held remotely.

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

April 5-9, 2021

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

Appendix 3: Panel membership

Virtual meeting through Google Meet, Point of Contact: Meaghan Bryan (AFSC/NMFS/NOAA)

Attendees:
Review Panel Chair: Kalei Shotwell (Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration or AFSC/NMFS/NOAA)

Senior Assessment Author: Meaghan D. Bryan (AFSC/NMFS/NOAA)

CIE reviewers:

- Sven Kupschus, Centre for Environment, Fisheries and Aquaculture Science (CEFAS), UK
- Colin Millar, International Council for the Exploration of the Sea (ICES), Denmark
- Anders Nielsen, DTU Aqua, Denmark

Other participants: all within programs at the AFSC/NMFS/NOAA

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