Sablefish
Stock Assessment Review (STAR) Panel Report

NOAA Fisheries, Northwest Fisheries Science Center
2725 Montlake Blvd. East
Seattle, WA 98112

July 8-12, 2019

Participants

Panel Members

John Field, National Marine Fisheries Service Southwest Fisheries Science Center (Chair)
Jim Ianelli, National Marine Fisheries Service Alaska Fisheries Science Center
Yong Chen, Center for Independent Experts
Robin Cook, Center for Independent Experts

Stock Assessment Team (STAT) Members

Melissa Haltuch, National Marine Fisheries Service, Northwest Fisheries Science Center
Kelli Johnson, National Marine Fisheries Service, Northwest Fisheries Science Center
Nick Tolimieri, National Marine Fisheries Service, Northwest Fisheries Science Center
Maia Kapur, School of Aquatic and Fishery Sciences, University of Washington
Claudio Castillo-Jordán, School of Aquatic and Fishery Sciences, University of Washington

STAR Panel Advisors

Patrick Mirick, Oregon Department of Fish and Wildlife, Groundfish Management Team representative
Gerry Richter, B&G Seafoods, Groundfish Advisory Subpanel representative
John DeVore, Pacific Fishery Management Council representative

Overview

The Stock Assessment Review (STAR) Panel met with the stock assessment team (STAT) in Seattle, at the Northwest Fisheries Science Center, from July 8th through 12th. This review focused on a single assessment, that of sablefish, throughout the duration of the week (commonly multiple assessments are covered at a STAR Panel). The Panel operated under the Pacific Fishery Management Council's (PFMC) Terms of Reference for Groundfish and Coastal Pelagic Species Stock Assessments (PFMC 2019).

The West Coast sablefish stock assessment was conducted using Stock Synthesis 3 (version 3.30.13), with the model period beginning in 1890 and ending in 2018. The model was based on the assumption of a single unit stock in the waters off of California, Oregon and Washington, although the STAT very clearly recognizes that the stock assumption was very likely violated, and STAT members are actively involved in efforts to evaluate tagging data, variable growth rates, genetic analyses and other information that will ideally lead to at least a research assessment throughout the range of the stock in the Northeast Pacific. The draft assessment included three coastwide fisheries, four surveys and an environmental index (relative sea level) that was modeled as a recruitment survey. Only one of those surveys, the West Coast Groundfish Bottom Trawl Survey (WCGBTS) is ongoing. The STAT expressed the greatest degree of confidence in estimating incoming recruitment and abundance trends from this survey.

Age and length composition data were initially included for all fisheries and surveys. However, due to conflicts between the age and length data with respect to growth and natural mortality, most length data were omitted. Selectivity curves were assumed to be age-based. Fisheries and surveys (except for the WCGBTS) were initially specified to be double-normal (e.g., dome shaped); and the final base case model allowed for dome-shaped selectivity for the WCGBTS as well. The model initially estimated length-based retention curves for discards based on available data. Natural mortality was estimated with a prior, while steepness was fixed at 0.7 and the data did not inform steepness. The length bin structure ranged from 18-90 cm (with 2 cm bin-widths) while the ages modeled were annual from 0 to 50 years for the data (the previous model went to 35) and 70 in the model dynamics. The compositional data initially used Dirichlet-multinomial data weighting but further explorations led to a data weighting that was most consistent with fitting the WCGBTS survey trend.

The Panel appreciated the model complexity and noted the length of the period modeled against the 71 ages in the dynamics highlighting the challenges associated with estimating growth and other factors. As such, model run times (including Hessian matrix estimation) ranged from 30 minutes to two hours which constrained the ability to handle requests promptly, particularly if a number of changes were suggested. Modeling the complex processes indicated by the data appears to be limiting the capacity of the analytical framework and software. Despite these concerns, the basic model result appeared robust to different model configurations, the trend

from the survey was informative, and most sensitivities and evaluations estimated the stock to be within the bounds of uncertainty for the base model with respect to depletion.

The STAR Panel recommended the sablefish stock assessment as the best available science, and that it provides a suitable basis for management decisions. Based on concerns raised during the review (highlighted in this report under technical deficiencies and unresolved problems and major uncertainties), the STAR panel recommended the next assessment be a benchmark assessment. However, if a benchmark assessment is infeasible in the near term, an update assessment may be warranted in the next assessment cycle due to the likelihood of recent strong year classes (2016, 2018) entering the fishery that are not yet well resolved in survey or fishery data. The panel noted that reductions in WCGBTS effort for 2019, and potentially beyond, may additionally challenge the ability to resolve recent year class strength estimates.

Summary of Data and Assessment Models

The STAT provided detailed presentations on available data and the main assessment approach. There were a number of clarification points and discussion that included issues related to age validation, reproductive capacity of older-age fish, catch estimation quality, and data quality. Catch estimates are based on a near census and appropriate conversions are applied from dressed to round-weight. Catch has been reconstructed several times with the same (similar) result.

The trawl survey time series provide “area-swept” indices but are fitted as relative for a number of reasons. A model-based approach has been adopted as standard for WC groundfish. It was noted that when a new year of data becomes available the entire index is re-estimated using the model-based approach, potentially allowing the time series to change. The extent of such changes might be tested retrospectively and, if substantial, may be indicative of schooling behavior or distribution changes or possibly movement changes in the north. The main survey duration is done in two passes from May-October every year.

Relative to model fitting, the analysts found that results are sensitive to data weighting. Discussions included how discard mortality was estimated (or set), and characteristics of selectivity patterns for different fisheries. While selectivity was modeled as aged based, it was noted that selectivity processes could be a result of some combination of size and age-based processes. There were discussions regarding how biological schedules (growth, maturity, fecundity) were estimated or specified, noting that they were assumed constant despite consensus that some (for example, growth or mean weight at age) vary over time due to environmental and spatial dynamics of the stock. Retrospective patterns appeared to be reasonable for both the pre- and post- review models, although the Panel noted that extending the retrospectives further back in time could help show the value of recent data.

In the period between the document distribution date and the review, the STAT made several notable changes to the model, including combining the two fixed gear fisheries (pot, hook and line), exploring alternatives to allowing for error in the sea level index, extending the length

discard dataset (new data from the West Coast Groundfish Observer Program, WCGOP), freeing the CV of young fish, and freeing various selectivity and retention parameters. The STAT and STAR explored these and many other changes throughout the course of the week, finding the model surprisingly sensitive to some changes (such as the addition of several years of discard length data), with scale (along with natural mortality) highly sensitive to a wide range of specifications and assumptions. However, many aspects of the key model results were fairly robust through most changes and sensitivity analyses. In nearly all cases the model estimated four very strong recruitment events over the past decade, fit the WCGBTS survey index reasonably well, and estimated an ending year biomass and associated depletion level that was around target levels. Notably, this was found to be the case even when the model was truncated to begin in 1970. The sea level index was included in the base model to inform recruitment, but was tuned in a manner consistent with how other survey data are tuned and consequently had only a limited influence on model results.

Requests by the STAR Panel and Responses by the STAT

The pre-STAR draft assessment document was very complete and the STAT's opening presentations to the panel were very comprehensive. Given that there was only a single assessment reviewed for this panel, time was used efficiently by receiving presentations on regional patterns of growth variability, the derivation of the sea level index, and other ecosystem considerations by other members of the STAT while the assessment leads addressed the first round of requests from the panel. Although some requests could not be filled in the time or sequence of the panel's original request, the STAT provided thorough responses to all important requests throughout the course of the week.

The requests by the panel are listed below based on the order and day of the request. Responses from the STAT team are given below each request, most of these responses were given the following day. Figures documenting many of the more significant results from the response to requests are also included.

Day one requests

Request 1: *Show the SS weight-at-age (over time) results and compare with NWFSC WCGBTS weight-at-age data, if possible.*

Rationale: The growth model is embedded in the assessment model and the variability (or lack thereof) may differ from the data. Also, to see if there's a temporal / year-effect pattern (e.g., due to strong year class(es) that may have cohort effect / density dependence).

Response: The STAT presented the requested plots (Figure 1). These plots show some systematic differences between the observed Northwest Fisheries Science Center

West Coast Groundfish Bottom Trawl Survey (WCGBTS) weight-at-age data and the SS internally-estimated weight-at-age data. The STAR Panel made a further request to evaluate the differences between the observed and estimated weight-at-age data, and the STAT produced Figure 1 (below). These plots show that these differences appear to be more obvious in recent years and there might be cohort effects. Such differences vary between female and male sablefish. The Panel identified the need to address this inconsistency, but decided to table this issue for now because more discussion is needed to identify potential approach to address this issue. The Panel considers this to be an important research topic to be considered in future assessments. Subsequent to this request, one member of the panel (Ianelli) conducted some additional data explorations to evaluate whether there were indications of cohort effects, which there is some suggestion of for female sablefish, but potential less so for males.

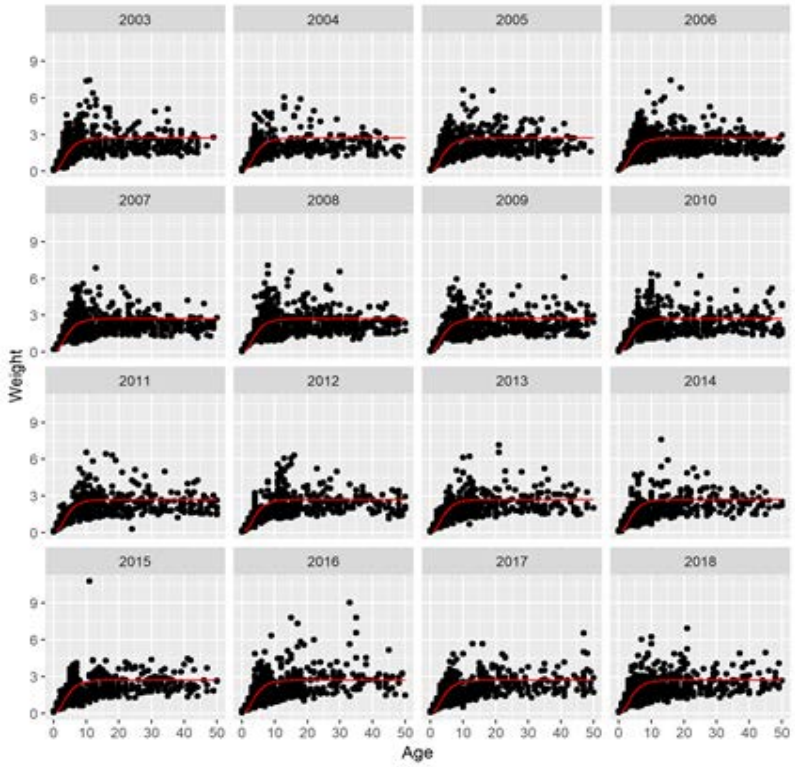

Figure 1: Model estimated weight-at-age over time relative to empirical (WCGBTS) weight-at-age data.

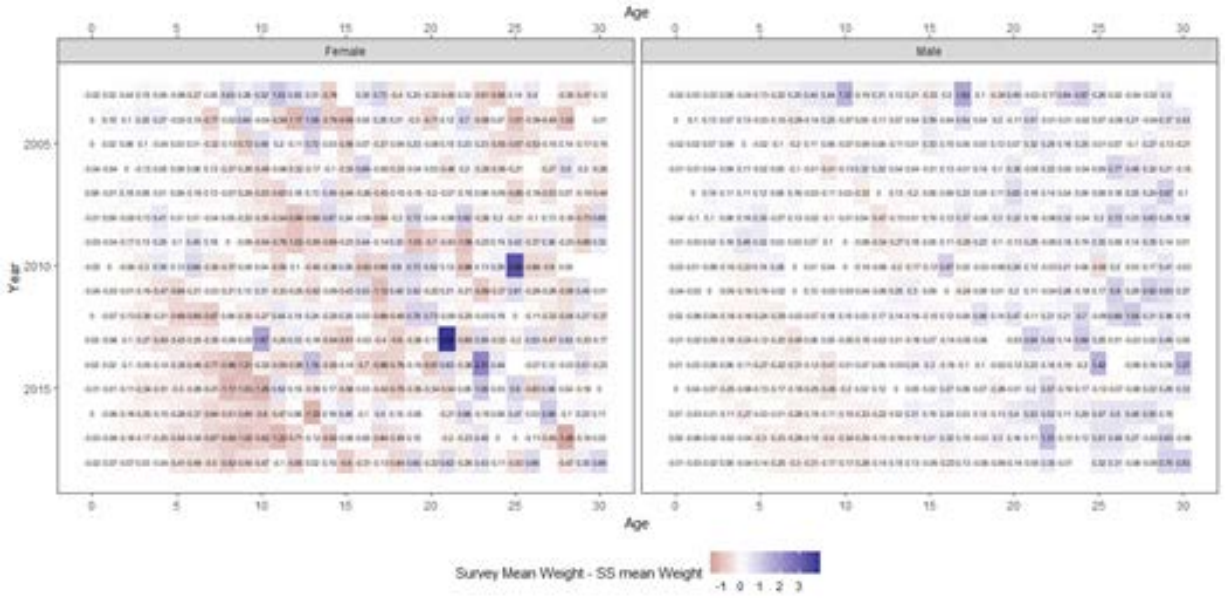

Figure 2: Empirical weight-at-age data and residual patterns from the WCGBTS survey for females (left panel) and males (right panel), indicating some potential cohort effects in mean size at age.

Request 2: *Plot cumulative size distribution for WCGBTS using the AKSLP survey footprint (N of 36^{o} and deeper than 100 fathoms), and compare with the AKSLP cumulative length frequencies (over all years).*

Rationale: The issue of setting the WCGBTS selectivity to be asymptotic is a change from past assessments and data supporting this specification, external to the model might be useful. Also, this may provide some justification for specifying asymptotic selectivity for the AKSLP survey data.

Response: The STAT provided the requested accumulative size composition plots for the three survey programs. The plots did not indicate substantive differences in the size composition of fish from the different surveys. After reviewing the size composition plots, the STAR Panel made an additional request for age-composition plots, which were provided by the STAT (Figure 3). The STAT and STAR Panel evaluated and compared the differences in size compositions and age compositions among the NW slope survey, AK slope survey, and WCGBT survey and found the differences are rather small (although slightly higher proportion of young fishes were found for the two slope surveys compared to that for the WCGBT survey). This may suggest similar selectivity patterns for the three survey programs. The STAR Panel recommends that further sensitivity analyses be done to evaluate model performance when all these three surveys have the same kind of selectivity curves (i.e., either

dome-shaped or asymptotic for all the three survey selectivity). Such patterns were explored in later requests to the STAT.

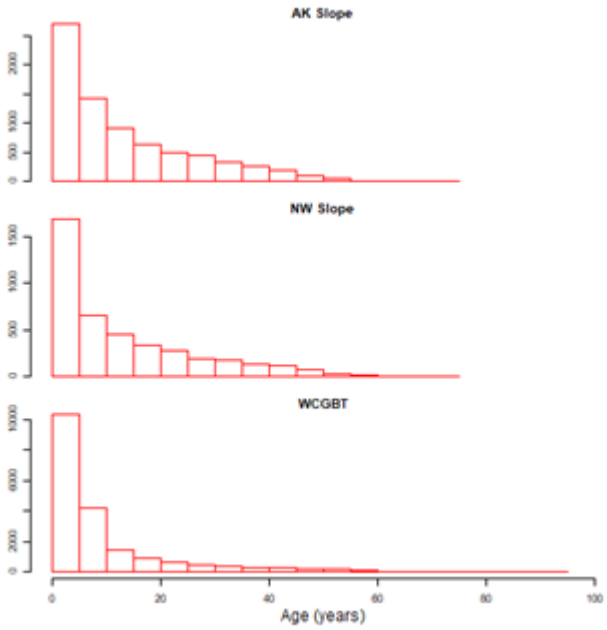

Figure 3: Raw age composition of three slope or shelf/slope surveys, all years combined.

Request 3: *Examine recruitment estimates from the base model and compute ratio of the sea level (SL) index to derive a q variability (CV) estimate (prior variance). Compare this with the assumed CV.*

Rationale: For the more informed period (e.g., 1980-2017 when survey are available), the recruits are based on age data and SL data may have little impact. The variability estimated from this period could be used for the prior for the variability used when recruitment data are less commonly available. This is to provide a more objective approach to specify the level of process error that might exist between SL and actual recruitment.

Response: The STAT consulted with Dr. Rick Methot who recommended adding an estimable error term to q, in a manner consistent with the variance inflation approach typically used for survey indices. This modification eliminates the need to make unnecessary assumptions on known prior variance for q. The STAT proposed that this approach was a more appropriate way to address tuning of the sea level index, and the STAR Panel agreed that this change improved the model parameterization and recommended that this new configuration be used to replace the q vector.

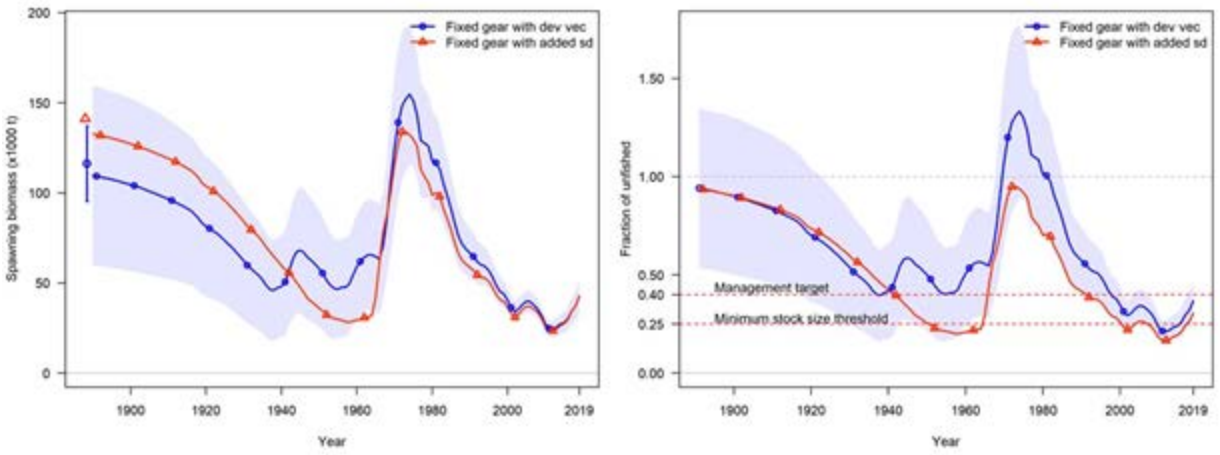

Figure 4: Comparison of the working base model with sea level tuning based on a deviation vector versus a single added variance parameter.

Request 4: *Starting with base-run model including WCGOP composition data (added post May 29), document July 1st and July 6th model changes, incrementally by characteristic (cumulative):*

- *Free young fish CV at age 0.5*
- *Free selectivity and retention parameters (the P6) and*
- *Include time-varying sea level catchability (SL q) deviation vector (and note assumed CV/prior)*
- *Combine HKL and POT fisheries into one. This reduces complexity and parameter estimation issues*

Include figures reflecting changes to each of these aspects. Specifically

- *How did CV change? Distributions of length at age in growth plots*
- *Selectivity curves changed*
- *Retention curves*
- *For SL q deviations, examine the time series of the values to evaluate variability in q*
- *Fits to length composition data to the new combined, HKL+POT fleet (residuals of combined compared to when split)*

Rationale: The STAT made these changes prior to meeting, and this will aid in understanding the impact of the changes.

Response: The STAT did all the requested analyses and modeling and provided the following plots (Figure 5). The STAT evaluated the changes in all of the likelihood functions and residuals in diagnosis plots induced by each change, and found that the incremental changes, combining with the addition of an error term in q (estimated to be 0.7 which was identified in the 3rd request) has improved the model fitting. The

panel noted that the model had unusually high sensitivity to the addition of four years of discard length composition data used to fit the retention curve for the fisheries. The STAT recommended that the model configuration with these changes made for the document base case be the "working base case" for further sensitivity analysis and evaluation. The STAR Panel agreed with the recommendation.

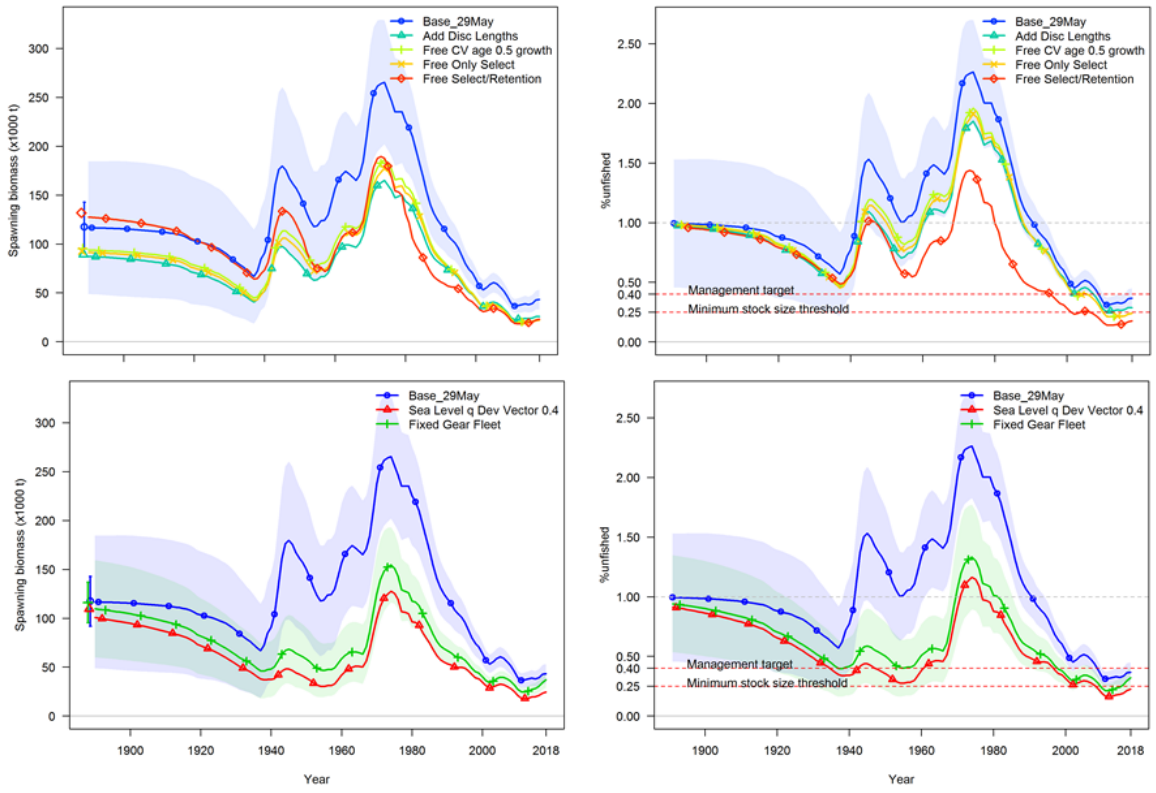

Figure 5: Sequential model results from STAR Panel request 4.

Request 5: *Test a case and just assume all bycatch/discards are dead rather than the current management values.*

Rationale: This issue arose in discussions of discard mortality estimates being poorly determined/estimated. This is just a sensitivity to highlight the relative importance of a field study to better estimate/revise given all the management changes (tow duration etc.).

Response: The STAT provided Figure 6 (below). The STAT and STAR Panel examined the changes and agreed that the observed result (a simple scaling upwards of biomass but no change to estimated depletion) is reasonable and easy to explain (i.e., need to have more fish in the population to account for additional discard mortality rates

assumed in this request). The Panel recommends that more studies are needed to better quantify the discard mortality for different fishing fleets, which can yield improved estimates of absolute SSB.

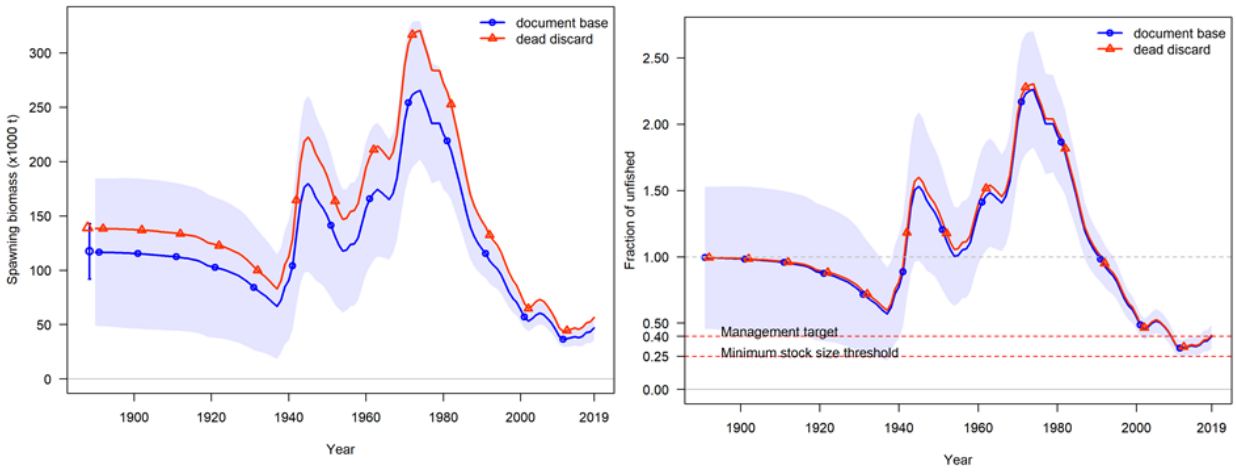

Figure 6: Working base model ("document base") relative to a model ("dead discard") in which all discards were assumed to have 100% mortality.

Request 6: *Test a run a shorter period. E.g., post 1970 and examine the B_0 and SSB trends.*

Rationale: This was intended as a way to evaluate the sensitivity on unfished spawning biomass in the absence of early age composition data.

Response: The STAT completed the request, and provided Figure 7 (below). The STAT and STAR Panel evaluated the changes in the SSB and depletion estimates for the model with a start year of 1970 compared to the model starting in 1890. The results indicated that the estimated unfished SSB and recent SSB values tend to depend on recent data. The choice of start years tends to only affect the stock dynamics between early and recent years, but has limited impacts on the estimates of SSB and depletion values at the start and end of the time series. Although this suggests that modeling time may not be necessary to start in 1890 for the west coast sablefish stock, the availability of reconstructed sablefish catch time series and common practice in the assessment of other groundfish stocks on the west coast makes the start time of 1890 more desirable. The STAR Panel agreed that the assessment start time should remain 1890 for sablefish.

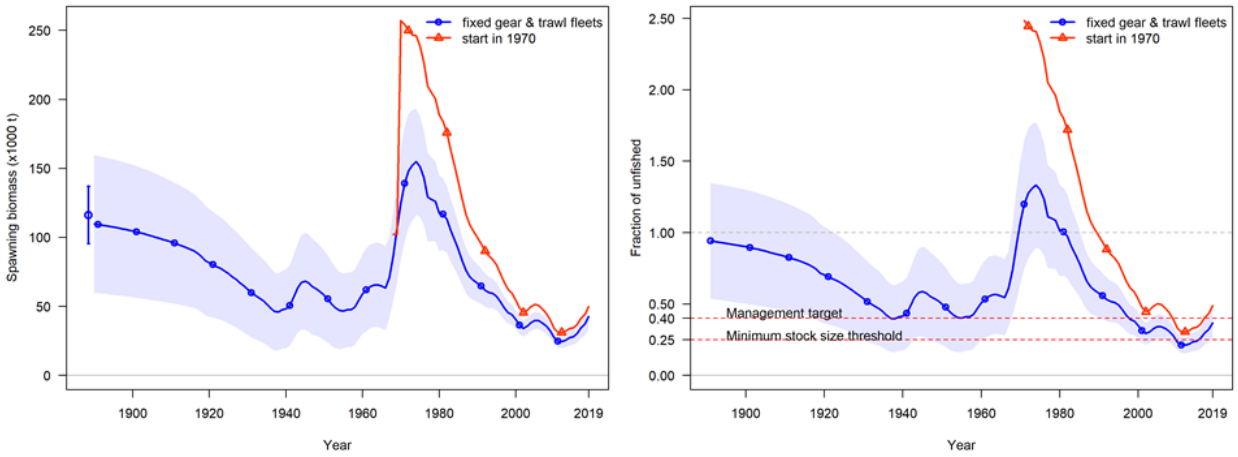

Figure 7. Comparison of the working base model with a model containing all of the same data, but that starts in 1970 with an initial F and initial biomass offset.

Request 7: *Examine a run omitting length data*

Rationale: The profile likelihoods seem to suggest that this data component differs substantially from most other components, there are a large number of size bins (with nearly no data) may be affecting the likelihoods, and when a sensitivity with "no age data" run was performed, the spawning biomass crashed

Response: This request could not be completed in the time allowed (was completed later).

Day 2 requests

Request 8: *Develop a selectivity sensitivity analyses, starting with the working base model agreed to this morning:*

- *Change the age-based selectivity curve to an asymptotic pattern for the NW slope and AK slope surveys.*
- *Leave the two age-based slope survey curves asymptotic and allow the WCGBTS to be domed shaped.*
- *Allow all age-based surveys to be domed shaped.*

Show model results as well as a comparison table and likelihoods across these alternatives. Split out the sea level index likelihood from the other surveys in these comparisons.

Rationale: When evaluating the length and age data from these surveys, these data were comparable among all surveys with some indication of proportionally older ages in the slope surveys.

Response: The STAT ran the model with three configurations:

a) all surveys having dome-shaped selectivity (dome dome dome)
b) dome selectivity for WCGBT but logistic for the slope surveys (log log dome)
c) logistic for the Alaska slope survey and the others domed (log dome dome)

The STAT also ran models in which all of these three surveys had logistic selectivity, and compared all four of these configurations to the working base model in which the two slope surveys had dome shaped selectivity and the WCGBTS had logistic selectivity (Figure 8). The run in which all surveys had dome shaped selectivity (run a, "dome dome dome" in Figure 8) had the lowest log likelihood, but it was more difficult to find a stable result for this run, and it was not possible to invert the Hessian matrix. This run also gave the most optimistic perception of the current stock status. All three runs gave similar estimates of R_0 and natural mortality (M), and estimated a similar unfished biomass. However, the biomass trend for run (a) diverged from about 1980 onwards compared to the other runs, including the provisional base run developed under Day 1 request #4, and implied a lesser degree of depletion.

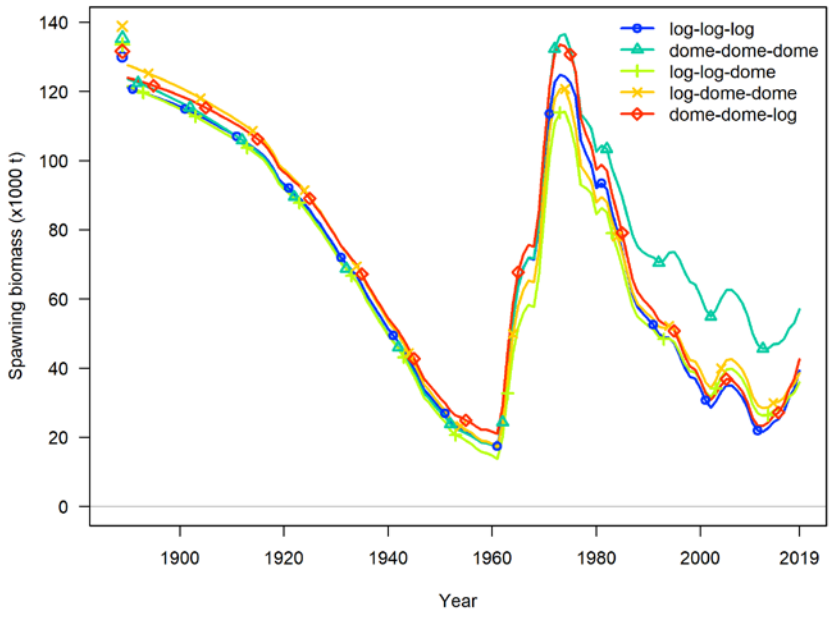

Figure 8: Spawning biomass estimates with a suite of logistic or dome-shaped age-based survey selectivity configurations for slope and shelf/slope surveys. Survey order is Alaska Center slope survey, NWFSC slope survey, NWFSC West Coast Groundfish Bottom Trawl Survey

As "run a" (all survey of these three surveys dome-shaped) appeared to fit the data better than the provisional base run, it was agreed that this should be pursued as the basis of a new working base model. However, there remained issues about the reliability of selectivity estimated for the fixed gear fleet. The STAT agreed to pursue a base model based on "run a" as described earlier (removing length composition data).

Request 9: *Re-run request 1c above with the length data removed (except for lengths in the discards). Show model results and the likelihoods in the comparisons requested in #1. If time allows, do a small number of jitters for requests 1c and 2.*

Rationale: The profile likelihoods seem to suggest that this data component differs substantially from most other components, there are a large number of size bins (with nearly no data) that may be affecting the likelihoods, and when a sensitivity with "no age data" run was performed, the spawning biomass crashed after the removals from the 1970s. It's possible that small errors in assumed constant growth curves affect length frequency predictions which may impact selectivity.

Response: This request could not be completed in the time allowed, largely as a result of time constraints in running models that were often unstable, but which estimated the hessian matrix to ensure convergence. Ultimately, model stability issues were resolved by fixing the CV of growth for young fish at the estimated value.

Request 10: *Provide a comparison of the working base model, with and without the sea level influence on recruitment, as well as with or without fishing (e.g., dynamic B_0 estimates). Provide model results, including comparison of recruitment and recruitment deviation estimates, and include a plot of the cumulative sum of the recruitment deviation vectors over time (not necessary for the dynamic B_0 runs). Also include a table of the changes in likelihood for the two runs with the sea level index specified in this comparison.*

Rationale: To understand the influence of sea level on recruitment over time, and to explore whether the cumulative values of recruitment deviation estimates indicate regime-like behavior in productivity.

Response: This request could not be completed in the time allowed.

Request 11: *Plot the recruitment values and deviations from the working base model without the sea level index and compare to the recruitment values and deviations in the 2015 assessment.*

Rationale: The 2015 relationship informed the sea level index used in the current working base model.

Response: This request could not be completed in the time allowed.

Request 12: *Provide two simple plots of growth estimates and mean lengths for ages 0 - 30 with factors being Regions (colors) for females and males with sex being on two panels.*

Rationale: To understand differences in growth within and outside the assessment area.

Response: The STAT provided Figure 9 (below), which suggests that using a single growth curve for data collected across the combined (west coast) regions may be inappropriate.

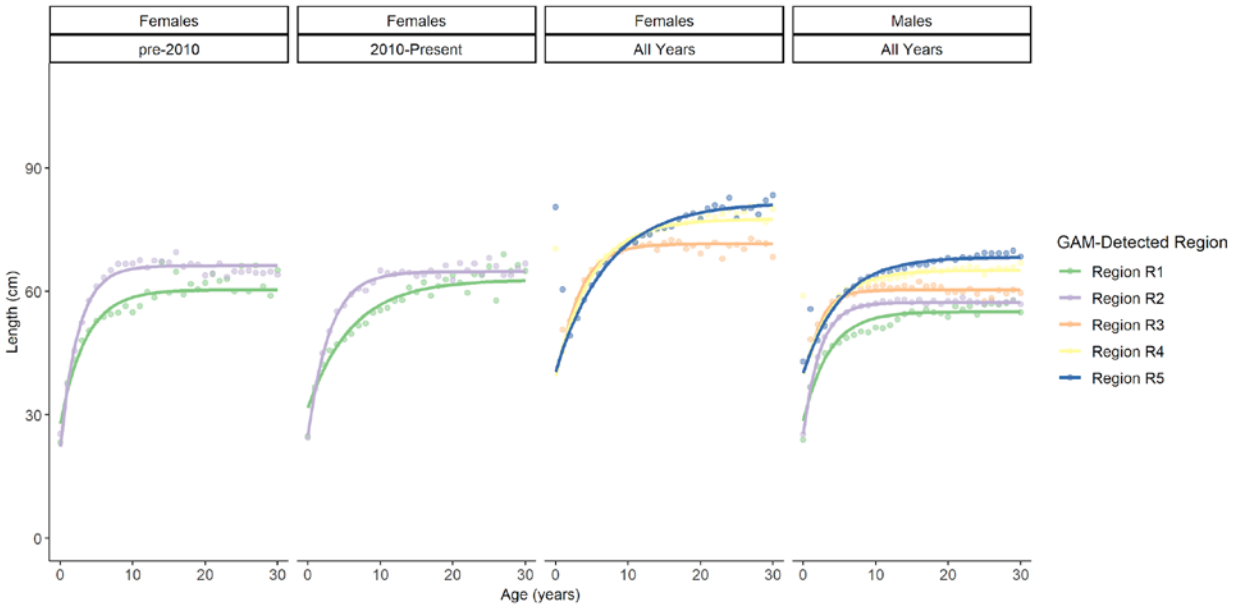

Figure 9. Summary results of sablefish growth curves in different regions of the Northeast Pacific. Region 1 is south of 36° N (southern California Current), region 2 is 36° N to 50° N (northern California Current), region 3 is 50 N to 130 W (northern British Columbia), region 4 is 130° W to 145° W (eastern Gulf of Alaska) and region 5 is west of 145° W (western Gulf of Alaska, Bering Sea, and Aleutian Islands).

Request 13: Using the working base model, do $F_{45\%}$ projections for 2021 and beyond catch assuming:

a. fixed gear catch only
b. trawl gear catch only
and display the relative catch values (or F_{SPR}s given equal catches) given a) and b) (no other model comparison needed).

Rationale: To understand that there may be some future variability in catch between actual gear types (irrespective of current "fishery" allocations), this will provide a baseline (extreme) range of the impact that will aid in future management considerations of future catch by gear scenarios.

Response: This request could not be completed in the time allowed.

Day 3 requests

Request 14: Attempt to get a model to converge with dome-shaped age-based selectivity for all surveys with the fixed gear fishery selectivity pattern estimated, if possible (for example, by constraining some parameters); otherwise, fixed at a reasonable pattern from a previous run. No sex-specific M for this run.

Rationale: The STAT explored a wide range of selectivity patterns and has not found an optimal model that converges. However, the STAT thinks additional effort towards this approach may lead to a base model. Additionally, there were discussions of evidence (net avoidance, larger and older sablefish encountered in hook and line surveys) that trawl survey selectivity may *a priori* not be expected to have asymptotic selectivity. Further, the prior distribution for sex-specific M did not suggest a difference in M and the data do not appear to be informative between the sexes.

Response: The STAT was able to get the two models (all age-based selectivity dome shaped for surveys, all surveys except WCGBTS dome-shaped) to converge by fixing the CV of growth of young fish (L_1 at A_1, fixed at the previously estimated values of 0.076 for females, 0.091 for males) as well as some of the parameters of the selectivity curve (all parameters for female selectivity to fixed gear were fixed, the males were estimated). The STAT also provided the results of runs that converged with the runs that had convergence issues on day 3, including comparison plots, a table of likelihood values, and key parameter estimates. The results were consistent with those reported for the model runs shown on day 3 that had convergence problems.

The panel asked whether the STAT had considered starting the growth curve at age 1.5 rather than age 0.5, where there may be less interactions between selectivity and size at age, as well as greater sample size. The question was inspired by some evaluation of size at age data for 0 and 1 year old fish from the WCGBTS explored by the STAR Panel on Wednesday afternoon. The point was made that the linear fit between age 0 fish and fish at the size associated with age 1 could be an undesirable property of such a model. However, the counterpoint, that it may actually be a desirable property, was also made.

The panel and the STAT also discussed the tradeoffs between fixing survey selectivities with fixing fishery selectivities, given that the model appears to need some selectivity patterns fixed to converge (the model had convergence problems when all dome shaped parameters were set to be freely estimated). The STAT proposed that total uncertainty was more robustly estimated when the survey selectivity was estimated rather than fixed. The shift in the perception of productivity is substantial between the runs in which all surveys are dome-shaped and in which all but the WCGBTS (and former AFSC triennial) are dome shaped. The model with all dome-shaped selectivities was more optimistic with respect to stock abundance

and productivity, although both models indicated that the stock has been overfished (SPR greater than the target level) in most years since the late 1970s.

Day 4 Requests

Request 15: Provide a run in which growth is estimated with conditional age-at-length (CAAL) data from the WCGBTS, the length data are removed from all fleets except for WCGBTS and the discards, and natural mortality is estimated as a single value for both sexes. Provide an additional run with the above changes in which the model begins in 1970 with an estimated initial F.

Rationale: Based on the results of the day 3 requests that were presented, there is tension in the age and length data influencing the growth curve. This may be a result of regional differences in growth that could interact with shifts in the distribution of fisheries effort, leading to greater tension in the model. The proposed base model is informed with age-based selectivities and the age data are thought to be the more important data to retain. Further, developing a model based on length data would require additional effort.

Response: The STAT provided the results of models with length data removed, with CAAL data used to inform growth, and with a single M. Not all of these were run with hessian matrices, given time available, so the STAT is not certain whether these models would converge. There were substantial changes in the perception of stock status in response to these runs, particularly in response to the run without length data (Figure 10). This suggests that the tension between length and age data are driving significant changes in the perception of stock status.

The natural mortality estimates among these three runs was variable, without the length data the estimate of natural mortality was considerably greater (0.083), with CAAL from the WCGBTS used to estimate growth, natural mortality (of females) was 0.048. As the model with "all" length data removed did not exclude discard length data, there was a fairly high likelihood component remaining in the "no length" model, indicating that the discard lengths are fairly influential. When the length data are removed, the age data pull the model towards a higher value of natural mortality.

The STAT provided likelihood profiles of natural mortality for the DDD model, which indicate that values between approximately 0.04 and 0.065 were plausible. The aggregated age data fit better with higher M values, the aggregated length data from the WCGBTS fit better with very low natural mortality estimates. The survey data fit better with a higher M. The Panel and STAT agreed that the age data should in principle be more informative with respect to natural mortality. Looking at likelihood profiles, all fleets except the WCGBTS fit age data better with higher M

values, the WCGBTS fit better with a very low natural mortality rate. The WCGBTS length data did not appear to be informative with respect to natural mortality, for all other fleets the length data fit better with lower M rates, with that effect being stronger for the fishery length composition data. The STAT noted that the 2015 model had similar patterns in the likelihood profiles.

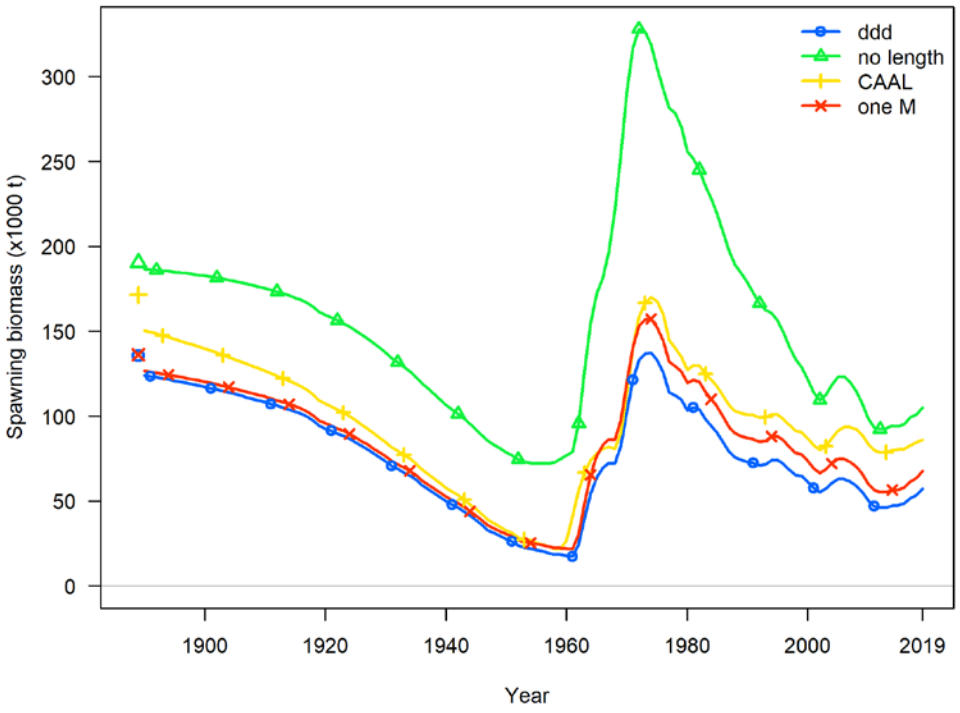

Figure 10: Results of the working base model from day 4, with (incrementally and sequentially), fisheries length data removed, CAAL data from the WCGBTS used to estimate growth, and with a single natural mortality rate parameter (e.g., no sex specific M).

Request 16: *Fix the retention curve for the discard length data for the fixed gear and trawl fisheries at their estimated values from the working base model and remove the compositional data from the likelihood estimation. Provide a comparison plot and table of likelihood and key parameter results for the two models.*

Rationale: These data are intended to estimate the retention curve rather than year class strength. As presently configured, the magnitude of the sample sizes from the discard lengths is substantial and may conflict with age composition data, which are more directly related to fishing mortality. This change should further simplify the model and reduce any remaining tension between length and age data.

Response: See response to Request 17, below.

Request 17: *With the working base model, try once more to estimate age-based selectivity for the fixed gear fishery. If a model that converges is found, provide a comparison plot and table of likelihood and key parameter results to this model relative to the models in the previous request.*

Rationale: To investigate whether reduced tension between age and length data may facilitate the estimation of the fixed gear fishery selectivity curve.

Response: The STAT provided plots of the results and a table of likelihoods and key parameter results. With the retention parameters fit there were a range of modest improvements and degradations to model fits to different data components, but no dramatic changes in likelihood to any one component. However, the spawning output and biomass trend scaled upwards significantly with this change, consistent with the previous observation that the discard length frequency data had an unexpectedly strong influence on scaling the model. In discussions, neither the STAT nor the STAR could explain why the discard length frequency data would have such a strong influence on scaling the model. There was agreement that the discard length data should not have an influence on scaling the overall model results. There was agreement that fixing the retention curve at the previously estimated values and removing the discard length data was appropriate for the working base model, and that future research should seek to evaluate why such discard data have an influence on scaling the overall abundance levels. There was discussion of the potential merits of exploring a methodology for estimating retention outside of the assessment model and fixing retention curves based on these external analyses in future assessments.

There was little change when (most of the) selectivity parameters for the fixed gear fishery were estimated, suggesting that recent changes have helped to stabilize the model. The STAT and STAR panel agreed that this was a reasonable improvement to the working base model.

Request 18: *Run a retrospective analysis.*

Rationale: Earlier runs suggested an unexpectedly strong influence of recent length data from discards. A retrospective analysis will help confirm that the model is not overly sensitive to recent data.

Response: There is a retrospective pattern of increasingly pessimistic perception of relative abundance as data are sequentially removed from the model. The potential cause of this was speculated to be the removal of the influence of recent strong year classes and higher abundance levels as inferred from the WCGBTS abundance trends.

***Request 19**: Do a run with the aging error turned off for ages 0-5. Provide a comparison with the previous base model result, likelihood values and key parameter values.*

Rationale: To ensure that the aging error is not influencing the ability to fit the age data for recent strong year classes, as there is an indication of under fitting in the age composition data.

Response: The STAT produced the results of the runs, which did not have substantial impacts on estimates of stock status but did slightly increase the relative strength of recent year classes. There were interesting patterns indicated in the overall compositional data fits, as this specification improved the fit to the age composition data (as expected) but at the expense of the fit to the survey data, and with a systematic mis-fitting of the aggregate age composition data at age 5-6. The abrupt shift in aging error estimates (e.g., from no error to substantial error) may not be appropriate. Greater investigation of aging error is recommended.

***Request 20:** Drop the last three years of sea level data (2016-2018). Provide a comparison with the previous base model result, likelihood values and key parameter values.*

Rationale: To ensure that these data are not drawing down the age and length composition data with respect to the strength of the 2016 year class.

Response: The results were provided and discussed, there was very little difference between the two runs. It was agreed not to remove recent sea level data from the model.

***Request 21:** Do likelihood profiles on the working base model, with any of the above changes that the STAT finds to be improvements, for ln(R0), M and steepness (in that rank priority).*

Rationale: To ensure no surprises in the current working base model.

Response: The STAT provided likelihood profiles of key model parameters. Results were comparable to previous evaluations. All age data except for that from the fixed gear fleet tended to want to inform a lower estimate of R_0 and lower estimate of M, length data (from the WCGBTS) wanted to estimate higher R_0 and M values. Survey likelihoods were not very informative with respect to R_0 or M, although it was noted elsewhere that higher M values improved the fit to the WCGBTS. None of the data were informative with respect to steepness.

Day 5 STAR Panel Requests

***Request 22**: Do a weighting sensitivity (Dirichlet multinomial, Francis, Harmonic Mean) and report the results.*

Rationale: To ensure that the model is insensitive to data weighting.

Response: The alternative weighting requests were completed and it was learned that the Dirichlet method resulted in less down weighting of the age composition and the single remaining WCGBTS length composition, as well as a poor pattern of residuals for recent survey index data. The Francis weighting method (similar to the harmonic mean approach) provided much improved fits to the survey index, was consistent with observed higher recent recruitments, and reduced the more extreme historical recruitments. The Francis and harmonic weighting approaches produced results that fell between the single and two sex natural mortality rate models that used the Dirichlet multinomial weightings. It was noted that the Dirichlet weights only adjust among length subsample sizes, such that if no other sample size weightings are possible (which is the case in this model with only WCGBTS length data), no adjustment can occur. The STAT and STAR panel agreed that the Francis weighting method (similar to the harmonic mean approach) was preferred, since this improved fits to the index data and more consistently weighted all the composition data (Figure 11).

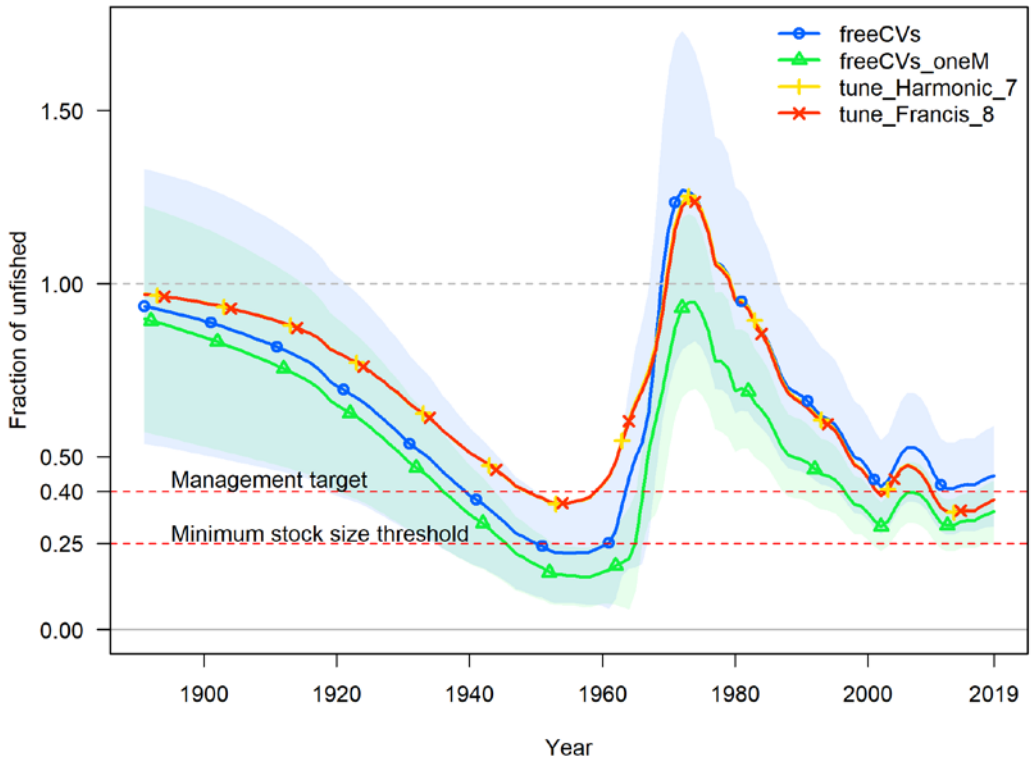

Figure 11: Projection of relative spawning biomass using different weight methods. The “free CVs” run was the then working base model with Dirichlet weighting, shown with both single and multiple M estimates, relative to tuned Francis and harmonic weighting.

Request 23:*Do a retrospective analysis of both the current working base and the single M sensitivity run. The STAT is free to report a subset of retrospective years (e.g., -2, -4).*

Rationale: The previous retrospective analysis did indicate retrospective patterns.

Response: Retrospective runs tended to revise biomass upwards as new data are added. The pattern was much reduced in the Francis weighing model compared to the Dirichlet-multinomial weighting, although there was still evidence of negative biomass in the estimates of biomass and relative depletion (Figure 12).

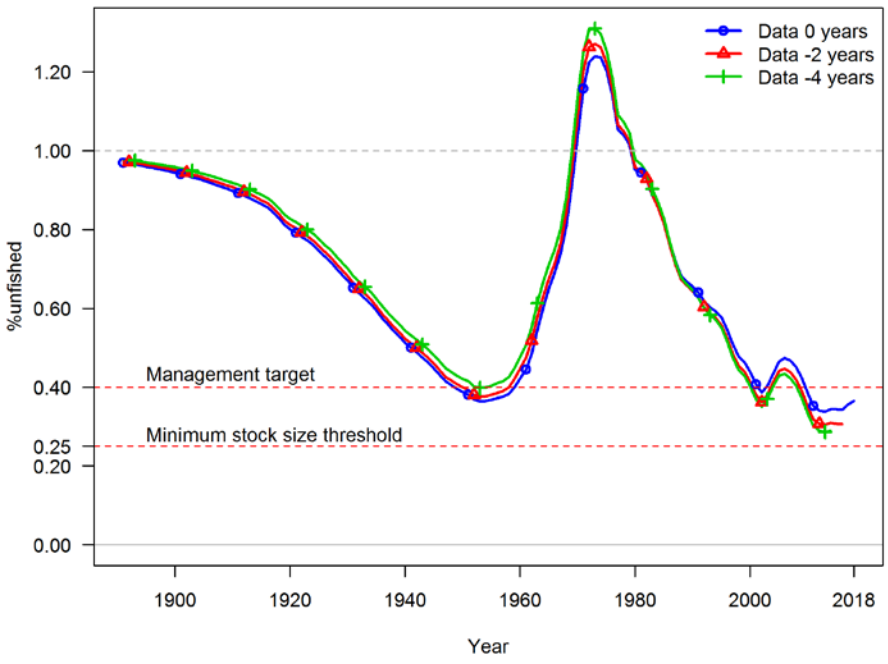

Figure 12. Retrospective (brief) runs from the accepted base model.

Request 24: Provide a first pass at a possible major axis of uncertainty for the decision table. Use the ln(R_0) point estimate that results in an ending spawning biomass consistent with the upper limit from the working base model, and the ln(R_0) associated with the ending spawning biomass from the lower 1.15 asymptotic confidence limit (e.g., 12.5% and 87.5% quantiles of 2019 spawning output estimates) for the single M sensitivity model.

Rationale: The terms of reference require an axis of uncertainty for the decision table.

Response: Due to the large number of day 5 requests, the STAT was unable to develop the decision table by the end of the meeting, but provided the draft decision table to the STAR panel the following week. The Panel found the decision table to be a reasonable approximation of the main axes of uncertainty for the base model.

Final model documentation requests

As the base model went through considerable changes over the week, there was insufficient time to do many of the sensitivity analyses that would have helped the panel understand the model dynamics and sensitivities, including changes in model structure that may have improved the model. Consequently, the panel strongly recommends that these analyses be included in the final model documentation, to help confirm that they do not substantially alter the fits to model data or the perception of stock status.

The Panel recommends the final documentation include the sensitivity to all three tuning methods, including the iterative results of Francis and Harmonic mean tuning, and a table of effective sample sizes over time. The original data-weighting approach proposed in the pre-Panel assessment document selected the Dirichlet-Multinomial. This showed promise for estimating weights dynamically (e.g., when doing profiles and retrospective analyses). However, following changes to the model prior to and during the review week, the resulting weights gave different results by method. Consequently, the group agreed that using the Francis method was preferred. As above, developing an assessment with time varying mean weight-at-age (empirically derived outside of the assessment model) should help resolve these (and other) issues. The final documentation should also explicitly state what methods were used to base starting effective sample sizes for each data source in the documentation (e.g., number of port samples, number of survey tows, etc.).

The Panel also recommends that the final document include preliminary analyses of the relative length and age composition data (frequency distribution plot, as well as a cumulative sum plot), of fixed and trawl gear fisheries north and south of 36° N for three periods, 1997-2002, 2003-2010, and 2011-2017 (periods corresponding to the influential management milestones provided in the draft assessment). The port of landing can be used as a proxy for region, such that the South of 36° N group would include Morro Bay, Santa Barbara, Los Angeles and San Diego port complexes, and the north of 36° N ports would start with the Monterey Port complex and all others further north. Ideally this will include a table of sample sizes (numbers of subsamples, number of length observations, number of age observations) by gear and region. To put sample sizes into context, a graph and/or table of total landings by gear type over time north and south of 36 ° N would be helpful.

The Panel recommends that the final documentation include a table or figure showing the predicted mean weight-at-age from the model compared with the observed mean values from the survey. Figure 2 of this report shows the time trend (from an earlier model) and this could also be updated with the latest estimates from the base model. The rationale is to highlight potential differences in the growth estimated within the model and those estimated from direct sampling and to note that besides apparent regional differences in growth, there is likely some year-to-year variability and perhaps cohort specific effects.

The Panel requested that a summary run omitting the WCGBTS index data as a sensitivity. This should assist the SSC and public on understanding the importance of the survey index.

Description of the Base and Alternative Models used to Bracket Uncertainty

After extensive discussions of results from the requests the STAT arrived at a model that could be considered as a base run with the following key characteristics:

1. Age-based selectivity for all four surveys (triennial shelf/slope, AFSC slope, NWFSC slope and WCGBTS) is allowed to be dome-shaped given data and other model configuration aspects.
2. The WCGBTS survey represents the most comprehensive source of information over space and time (greatest range of depths covered, entire U.S. coast covered, longest of the time series), and thus should provide the best information regarding stock status.
3. The two fisheries in the revised base model (trawl and fixed gear) also use dome-shaped, age-based selectivity curves, with offsets for males relative to females, and time blocks that account for major fisheries or regulatory regime changes.
4. The model uses the WCGBTS age and length composition data to inform the growth relationship using CAAL data.
5. The model omits the length compositions from the other sources since age data are available for most of the same years.
6. The retention curve parameters were estimated including the discard length composition data within the model and then fixed at those values in the final model (and ignoring the discard length composition data as it was affecting the scale of the population in relative terms in ways that appeared to be inappropriate).
7. Key parameters of the fixed gear selectivity are estimated
8. The Francis weighting method (similar to the harmonic mean approach) is used since this improved fits to the index data and more consistently weighted all the composition data
9. The environmental index (sea level) is included but, as with other indices, an additional variance term was estimated.
10. Steepness is fixed at 0.70, as it was in the draft model, while natural mortality is estimated, separately for males and females, with informative priors.

For the uncertainty application, the group agreed that using the profile likelihood of lnR0 to bracket the range between the high and low states of nature, based on the asymptotic confidence limits, was appropriate. This was done after the STAR panel concluded.

Technical Merits of the Assessment

The WCGBT survey data appear informative with respect to both incoming recruitment and population trends. The assessment incorporates the latest information and applies growth estimates based on fitting the conditional age-at-length (CAAL) data from the WCGBTS.

There seems to be far less uncertainty regarding historical catch data than there is for many stocks that were historically landed in mixed-stock market categories. There is adequate fisheries composition data to inform the demographics of fisheries removals.

Despite several concerns raised by the STAR panel, the basic model result appears robust, the trend from the survey is informative, and most evaluations estimate the stock to be within the bounds of uncertainty for the base model with respect to depletion. The Panel and STAT agreed that the current base model is an improvement over the 2015 assessment model and can be used for management advice.

The STAR panel recommends that this assessment be considered a tier 1b assessment, as there are reliable age composition data sufficient to resolve year class strength and growth, and there is an information on trends from a fishery independent survey (WCGBTS). The panel recommends that the next assessment be a benchmark, due to the technical issues highlighted below and throughout the review. Barring doing a full benchmark, an update in 2021 should be done as there are signs of recent strong year classes that are not yet fully realized in recent trawl surveys.

Technical Deficiencies of the Assessment

There were odd behaviors on how critical estimates (e.g., of natural mortality) changed during the week of review. Specifically, there was little indication that the data were informative with respect to differences in natural mortality rates between the sexes, as there was not always consistency in which gender had the higher natural mortality rate.

There is a mismatch between the level of growth variability allowed within the model and the observed empirical data (specifically, the mean-weights at age). Given the large number of years and ages presently included in the model, developing a model with an approach that estimates the appropriate level growth variability will increase computation time dramatically, likely confound more parameters (cause estimation issues), and make reliable uncertainty estimates more difficult. A reasonable interim approach would be to develop empirical weight-at-age matrices that captures observed trends in growth (and speed computations).

The approach used to estimate the retention affecting the discards within the model in the first pass then fixing the values and removing the influence of the discard length composition data was an approximation (including the length data significantly affected model results in

unexpected ways unrelated to the shape of the retention curve). This practice was identified as unique and the behavior and interaction with when the data are included was unclear and should be investigated further (e.g., by changing the likelihood function and data input to something relating more directly to the retention curve rather than as "composition" data that affects recruitment estimates etc.).

Areas of Disagreement Regarding STAR Panel Recommendations

There were no major areas of disagreement among STAR Panel members (including GAP, GMT, and PFMC representatives), nor between the STAR Panel and the STAT.

Management, Data, or Fishery Issues raised by the GMT or GAP Representatives During the STAR Panel Meeting

The GMT representative noted that while the sablefish stock is now projected to be healthy with the 2019 base model, the future OFLs will be lower than those of the 2015 assessment due to reductions in the scale of the spawning biomass. The STAR panel noted that scale was not well informed in this assessment as evident across alternative model runs that had similar likelihoods at high and low spawning biomass levels. To better inform scale in the future, the GMT representative and STAT both recommended that representative estimates of absolute abundance be developed from the bottom trawl survey. Expansions would have to be made for fish missed by the survey such as the older fish that can outswim the trawl or for fish that occur deeper than the survey footprint.

The GMT representative also recommended that separate area models be considered in the future to better inform harvest specifications for the northern and southern sablefish management areas (36° N. lat.). Currently, the coastwide ABC from the single coastwide assessment is apportioned to set ACLs for the northern and southern areas based on the estimates of abundance from the trawl survey. Separate area models could better inform the northern and southern ACLs because there are considerable differences in growth between the two areas, often large differences in recruitment, and potentially differences in exploitation rates. At the minimum, having separate fleets-by-areas could help resolve some of the issues associated with differences in growth.

A member of the public (Mike Okoniewski) discussed the issues relating to sablefish is a bycatch component for other important fisheries. While ex-vessel value is an important metric to measure socio-economic outcomes, the role sablefish occupies as bycatch for the success of other fisheries is also important. While not expansive, Appendix A of the draft assessment has captured many of these concerns. Consideration of all these criteria should be integrated holistically into the context and process of our management decisions and policy choices.

Unresolved Problems and Major Uncertainties

There are indications of spatial differences in growth that appear to be related to tensions between age and length composition data in the model, particularly as related to commercial fisheries (trawl and fixed gear). Maps of relative effort in fisheries included in the assessment are also indicative of substantial shifts in effort over time, with declines in trawl fishery effort south of Cape Mendocino following the implementation of the catch share program in 2011, and increases in fixed gear (hook and line, pot) effort and catches south of 36° N in the post-2011 (much of which is a consequence of gear switching by catch shareholders). Greater consideration of the potential to improve the model by splitting fleets into different areas may benefit future assessment efforts.

There is a mismatch between the level of growth variability allowed within the model and the observed empirical data (specifically, the mean-weights at age). Given the large number of years and ages presently included in the model, developing a model with an approach that estimates the appropriate level growth variability will increase computation time dramatically, likely confound more parameters (cause estimation issues), and make reliable uncertainty estimates more difficult. A reasonable interim approach would be to develop empirical weight-at-age matrices that captures observed trends in growth (and speed computations).

The approach used to estimate the retention affecting the discards within the model in the first pass then fixing the values and removing the influence of the discard length composition data was an approximation (including the length data significantly affected model results in unexpected ways unrelated to the shape of the retention curve). This practice was identified as unique and the behavior and interaction with when the data are included was unclear and should be investigated further (e.g., by changing the likelihood function and data input to something relating more directly to the retention curve rather than as “composition” data that affects recruitment or other model estimates.

The indices are based on annual index estimates from the VAST model or, for the environmental sea level series, using dynamic factor analysis. It was unclear if the terms and application of these models provide index estimates that are independent and identically distributed (IID) by year. Since SS3 treats the index values as such, the covariance term over years should be examined as a check. If the correlations over time is significant from these index models, then the values submitted to the assessment model (based on the diagonal) would be inappropriate.

Recommendations for Future Research and Data Collection

Short term

To understand whether regional differences in growth, and associated size or age composition, are behind the strong tensions observed in age and length composition data, future assessment

models should evaluate having separate fixed gear fleets north and south of 36° N (as discussed in unresolved problems and major uncertainties). The potential merits of other regional differences in fleet structure could also be explored.

Future assessments should consider the use of empirical weights-at-age. One approach would be to begin with the WCGBT data and use it to fit (outside model) a smoother outside the model (see Jim Ianelli for code that uses a random effects model to estimate weights using empirical data by cohort and year). This would avoids confounding weight at age with growth estimates (e.g., Lee et al. 2019, Whitten et al. 2013), should allow the model to run faster, and should enable the model to better accommodate variability in growth (in both age and over time). Another advantage is that somatic body masses are based on actual measurements instead of model estimates which predicts mean length and then converts length to mass via a fixed set of length-weight parameters.

Consider (and potentially adopt) length based selectivity for the WCGBTS, as selectivity bias could lead to observed age zero fish not being a good representation of mean length at age 0 (which could affect estimates of growth). Figure 13 (below) shows some summary results of this issue, including the lack of apparent growth of age 0 sablefish during the period of the survey. A related issue is that specifying the size at age 0.5 (or 1.5), as well as the CV, should be in common by sex. Dimorphic growth presumably occurs at older ages. Future assessments should consider configuring "offsets" for sex differences in the growth. The rationale for this change is that differences between sexes at such young ages seem unlikely and re-configuring the model in this way will reduce the number of parameters needed.

Examine age-specific selectivity that is fixed at a "plus group" age (or have a reasonable rationale as to why selectivity might be changing over those ages).

Provide estimates of retention external to the model and evaluate if the approach of doing it within the model (and then fixing the values because of discard length composition data affecting model results other than the shape of the retention curve).

Reconsider spline or some other non-parametric selectivity forms, as there are unexpected behaviors observed in estimating double normal selectivities for the fixed gear fishery. The requirement (apparently) to have to fix the "P6" parameter related to old sablefish selectivity and some other interactions seems unusual, and several of the convergence issues were related to selectivity parameters for some gear types/surveys. The 2015 model used a different form for selectivity and the document provided little justification to choose one over the other. "Fewer parameters" is a poor reason if performance overall is so much worse. The convergence problems did seem to be largely resolved after the STAT fixed the CV of young ages.

The CAAL plots suggest that in the early years of the WCGBTS the residuals of the biggest fish were generally "young" and in the more recent years the pattern of observed and predicted were

more consistent. This could be due to changes in growth and/or size (or age) based selectivity/availability changes and should be investigated.

Developing diagnostics that better evaluate sex ratio observations against model predictions would be useful, especially given some of the differences observed in the fisheries selectivities estimates and in sex specific natural mortality configurations compared to combined sexes natural mortality.

Reducing the number of ages should be considered, as the plus group information and dynamics beyond some ages are unlikely to change (as shown in Figure 14).

Continuing to evaluate the use of sea level and/or other environmental indicators as drivers of both historic and future recruitments is strongly encouraged, and future assessments should strive to clearly convey how such data explicitly relate to recruitment as well as to convey the logic and presumed mechanisms behind the relationship.

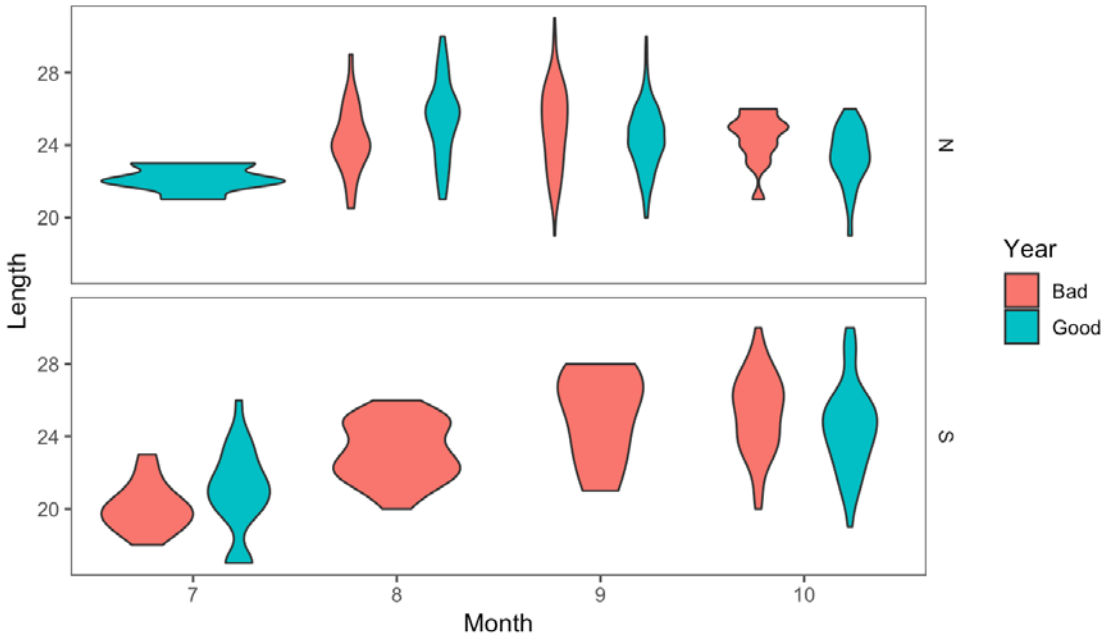

Figure 13. Distribution of WCGBTS length (vertical scale) for sablefish at age zero (2003-2018) by month, area (N=north of 36° N, S=south of 36° N; panels) in “good” years and “bad” years of recruitment. “Good” years are 2008, 2010, 2013, and 2016 while the “bad” are the other years.

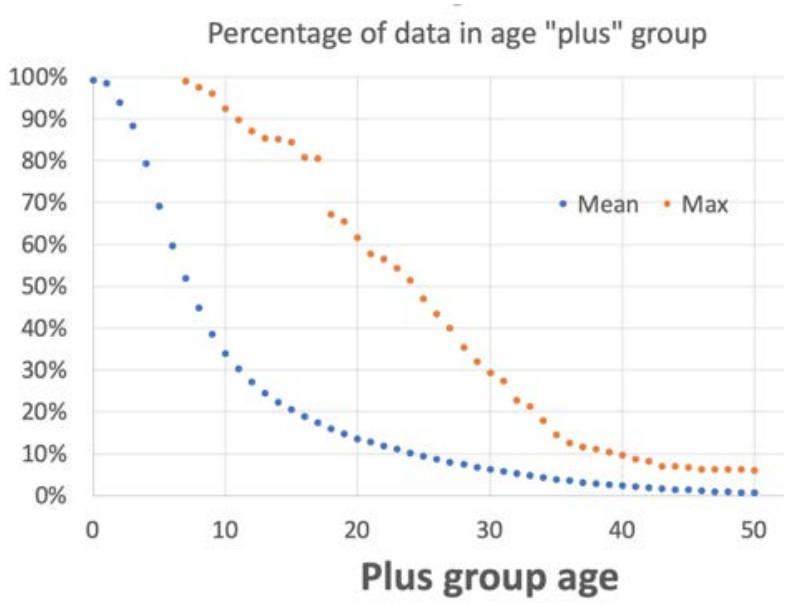

Figure 13. Profile showing what percentage of sablefish are in “plus” group based on input data. The “mean” represents the mean percentages over all available age data used in the assessment and “max” represents the maximum percentage over all data types and years.

Medium and Long Term

The panel had concerns about the large amount of data and the challenges faced doing things such as estimating natural mortality and the CVs on growth and other factors. However, the core growth parameters were generally stable, and convergence problems are not atypical when natural mortality is difficult to estimate and selectivity patterns are all dome-shaped. The panel notes that it is worth reflecting on whether a somewhat different modeling approach would work better (in the longer term). The nature of the issues in the data may be stretching the capacity of the analytical framework and software to do what needs to be done.

The STAT provided very helpful updates on ongoing efforts to evaluate life history characteristics, movement patterns and management approaches throughout the range of sablefish in the Northeast Pacific. The STAR Panel agrees with the STAT that ongoing and future work, such as efforts to develop a transboundary stock assessment and management framework, be pursued. This is based on strong indications that current stock boundaries are likely to be inappropriate, and that a transboundary assessment would likely improve the ability to estimate the scale of the population.

As the WCGBTS is highly informative in the model, maintaining full coastwide survey effort is essential. However, currently the survey does not include a large fraction of the habitat south of 36°N, the cowcod conservation areas. Despite a lack of data in this large area, catch is allocated

north and south of 36°N based on the estimated fraction of sablefish in these areas, and this fraction, in turn, is based on an extrapolation of survey catch rates outside the CCAs to those inside. As fish within the CCAs are only subject to fishing pressure if and when they move, and movement rates are variable, this concentration of effort outside of the CCAs could potentially lead to localized depletion, which in turn could bias the signal in fishery (and potentially survey) age and length composition from the fished areas. It would be beneficial to have survey data from within the CCAs to inform the survey, and to allow for some evaluation of whether and how population structure may vary inside and outside of the CCAs. There could be some potential for local depletion elsewhere as well, given the concentration of trawl effort off of Oregon and Washington and the decline in fishing effort and catches of both trawl and fished gear in California north of 36°N.

For the WCGBTS, evaluate cohort total mortality (Z) for consistency and as a check with model values. Comparing the survey estimate numbers at age over time (e.g., relative abundance at age 2 in year y compared to abundance at age 3 in year $y+1$).

Better estimates of aging error, bias and continued efforts to improve on age validation remain high research priorities.

Acknowledgements

The panel thanks the STAT for their hard work, openness and responsiveness during the review. The panel acknowledges that as this panel focused on a single assessment (rather than two or more models, which is more typical of a star panel), that the depth of review and workload associated with filling panel requests was considerably greater than typical levels. The panel also thanks Stacey Miller, Jim Hastie and others at the NWFSC for their hospitality during the review.

References

Lee, H., Piner, K.R., Taylor, I.G. and Kitakado, T., 2019. On the use of conditional age at length data as a likelihood component in integrated population dynamics models. *Fisheries Research* 216: 204-211.\

Pacific Fishery Management Council (PFMC). 2019. Terms of reference for the groundfish and coastal pelagic species stock assessment review process for 2019-2020. https://www.pcouncil.org/wp-content/uploads/2019/04/Stock_Assessment_ToR_REVISED_2019-20_APR2019_Final-2.pdf

Whitten, A.R., Klaer, N.L., Tuck, G.N. and Day, R.W., 2013. Accounting for cohort-specific variable growth in fisheries stock assessments: a case study from south-eastern Australia. *Fisheries Research* 142: 27-36.