Agenda Item E.6.a
Attachment 2
June 2007

Sablefish STAR Panel Report

National Marine Fisheries Service
Hatfield Marine Science Center
Captain R. Barry Fisher Building
2032 S.E. Oregon State University Drive
Newport, Oregon 97365

May 7-11, 2007

Reviewers:
Martin Dorn, Scientific and Statistical Committee (SSC) Representative, STAR Panel Chair
Patrick Cordue, Center for Independent Experts (CIE, Rapporteur)
Vivian Haist, Center for Independent Experts (CIE)

Advisors:
Heather Mann, Groundfish Advisory Subpanel (GAP) Representative
Mark Saelens, Groundfish Management Team (GMT) Representative

Stock Assessment Team:
Michael Schirripa, Northwest Fisheries Science Center (NWFSC)

Overview

The West Coast sablefish stock assessment assumes a unit stock in the waters off Oregon, Washington, and California. The draft stock assessment was conducted using a recent version of Stock Synthesis 2 (SS2) and used data from many sources. Five alternative model configurations were presented, with the proposed base model using nine data sources including two environmental variables as recruitment indices. The proposed configurations excluded logbook and pot survey indices which were used in the previous assessment. All model configurations included available length, age, and biomass data from four bottom trawl surveys of the slope or shelf. Available length and age data from trawl, hook and line, and pot fleets were included. The estimated catch history extended back to 1915 (split into the three gears). All proposed configurations had the proportionality constant (q) for the NWFSC slope survey fixed at 1.

The STAR Panel was concerned about many aspects of the proposed base model. In general terms, there were three main concerns. First, the model appeared too complex relative to the expected information content of the available data. Second, the age, length, and length-at-age data from each data source were input into the model as if they were independent (when they clearly were not). Finally, the assumption of q=1 had no firm basis. In specific terms, the STAR Panel had one over-riding concern with the base model. The model did not fit the NWFSC slope survey abundance indices despite the assumption of q=1. The expected values were almost all larger than the observed values (they went “over the top”).

The STAR Panel and STAT worked towards a new base model by making progressive changes to the proposed base model. Minor changes included the incorporation of some discard rate data that were available but had not been used, a tightening of the allowed variability on the annual fishery selectivities, and the exclusion of the zooplankton time series. Iterative re-weighting procedures were also applied to the age and length frequency data sets.

An important change was made to the NWFSC slope survey biomass indices. In some years, in the Conception stratum, all trawl stations were north of Point Conception and the average catch rate had been applied to the whole stratum area – despite catch rates being much lower south of Point Conception. A new biomass time series was obtained in which the Conception stratum extended only to Point Conception.

An informative prior was developed for the NWFSC slope survey q by considering its individual components and using the expert opinion of meeting participants (and some data) to bound each component. The median of the prior was obtained by using the “best guesses” for each component. The range on q was (0.22, 0.86) with a “best guess” of 0.56. A prior was formed by equating the “bounds” to 99% of the distribution. The prior was used in a model run but the estimated q was well outside the “bounds.” The decision was made to fix q at the median of the prior for a revised base model.

After the above changes were incorporated the revised base model still exhibited the "over the top" problem. This problem was addressed by down-weighting the commercial fishery age and length frequencies (by shifting the emphasis level from 1.0 to 0.1). This is a pragmatic approach, which the STAR Panel and STAT agreed was justified given the uneven spatial and temporal coverage of the commercial fishery sampling (and hence the large potential that the data were not representative).

Uncertainty in the base model was represented by three sensitivity runs: a lower *q*, a higher *q*, and a run excluding the environmental time series.

Analyses requested by the STAR Panel

<u>Round 1 requests</u>

A: Use discard rates from Pikitch et al. (1988) and the ADCP database and interpolate where necessary. Briefly compare model results to base-run model (spawning biomass trajectory).

B: Two requests re the NWFSC slope survey:

1. Produce a plot of the proportion of biomass in the Conception stratum each year.
2. Obtain mean catch rates and biomass estimates for north and south of Point Conception for each year.

C: Time series of selectivities for each fishery:

1. Plots of selectivity at length, for the following lengths:

H&L:	51 cm
Pot:	49 cm
Trawl:	45 cm

2. Plots of selectivity at age 40 (all fisheries).

D: Develop an age-only run with the following specifications:

1. Fit only to biomass indices and age frequencies.
Three fisheries.
Estimate age-based selectivities (constant for each fishery)
Estimate recruitment deviations.
Estimate all *q*s
Fix steepness = 0.5

2. Variation on D1: allow selectivities to vary annually to better fit age frequencies

3. Variation on D1: *q*=1

Round 1 responses

Run A showed little change from the original base model (Configuration 4).

A request was placed to obtain the NWFSC trawl survey data and biomass estimates – reporting on Request B was delayed until the data arrived. Further requests were made with regard to this biomass time series which superseded/included this request (see Round 2 requests).

The time series of length and age selectivities were produced and showed very wide variation in selectivity across years. Few consistent trends, if any, were visible. The suggestion was made to tighten the sd on annual deviations (see Round 2 requests).

All three runs in request D were completed and the diagnostics for run D1 were examined in detail by the Panel and STAT. Several issues arose during the presentation of the results.

The poor fit to NWFSC slope survey biomass indices seen in the base model was fixed but the estimated q was unrealistically low. A poor fit was noted for NWFSC slope survey age frequencies in the plus group, with too many males observed and not enough females (relative to the models expectations). Peculiar data in the age-length observations for AFSC slope survey were noted – the data at times appeared too regular, forming perfect linear relationships. Some peculiar fits to age data were also seen – there appeared to be very high predicted values for some young age classes, which implied a strong cohort within the model, but when the age-0 recruits were examined, the cohorts were not strong (Dr Haist suggested that the predicted values must be accumulations from age 1 – and she was subsequently proven correct – an SS2 feature.)

The biomass time series for NWFSC slope survey was again discussed – in particular the change from a GLM approach to an area-swept approach. The STAT expressed concern about the GLM results used previously and the interpretation of them as biomass. Hence the STAT argued in favor of a change to the simpler and more easily understood swept-area method. The iterative re-weighting method used by the STAT was discussed – the method used was determined to be inappropriate as it potentially changed the relative sample sizes across years within a data set.

Runs D2 and D3 were briefly examined. The estimation of annual selectivities in D2 gave a much improved fit (200 likelihood units, for 60 constrained parameters), but annual deviations were still “very wild”. Run D3, which again fixed *q*=1, gave similar results to the base model and exhibited the same bad fit to the NWFSC slope survey biomass indices (going “over the top” of the observations).

There was a verbal request for the STAT to try a variation on D3 with R_0 fixed at 80% of its estimated value. The idea being that the model would then be forced to fit NWFSC slope survey biomass and there was interest in seeing which likelihood components degraded (and hence which data sets "preferred" the higher biomass – and hence caused the "over the top" problem.) This run did not initiate successfully. A "crash" penalty occurred almost immediately, a consequence of exceeding the maximum exploitation rate. Dr Haist requested that the input files be made available to her to investigate. The STAT volunteered to investigate the cause of the "over the top" problem using the base model.

Round 2 requests

E: Two runs building on the "progressive" base case (see A in Round 1 requests):

A1: Change sd on annual deviations for selectivities to 0.35; and use the recommended iterative re-weighting procedure

A2: In addition to A1, change to "north of Point Conception" biomass time series.

F: Produce a plot of the "north of Point Conception" biomass time series from swept area and the corresponding GLM time series used in the previous assessment.

G: Investigate and report on the strong yet average cohort seen in run D1 (strong on age fit but not a strong cohort (in 1988 and 1999 age composition) according to age 0 recruits).

H: Investigate and report on the "too regular" age-length data.

I: Plot the biomass estimates and/or density (kg/ha) north and south of Point Conception within the Conception INPFC area.

Round 2 responses

Runs A1 and A2 were partially completed because the iterative re-weighting was not repeated long enough for convergence. The changes made little difference to the results and the "over the top" problem still existed. When *q* was estimated in a variation of A2 the estimate was again unrealistically low.

The new area-swept "north of Point Conception" time series and the corresponding 2005 GLM time series showed similar trends but with the GLM time series at a higher absolute level (and curiously expressed in units of "1000 t / 2 ha").

Requests G and H were not able to be done (request G was subsequently not needed as Dr Haist's suspicion was later confirmed; and request H was repeated in Round 3 requests).

The density estimates north of Point Conception within the Conception stratum were shown to be typically much higher than those south of Point Conception (confirming that the NWFSC time series used in the original base model was inappropriate).

Formation of a prior for the NWFSC slope trawl survey

The STAR Panel Chair suggested to the STATs (for both sablefish and longnose skate) that it could be beneficial to construct informed priors for the trawl surveys where each of them had fixed *q*=1. A joint session was held for this since the proposed method was identical for both species. There was a general discussion on what the approach entailed and both STATs agreed to participate. The general approach described below has been used in New Zealand for several years in one form or another.

The approach requires that the trawl survey *q* is split into three components: areal availability (the proportion of stock biomass in the trawl survey area), vertical availability (the proportion of biomass in the water column that is available to the trawl after vertical herding), and vulnerability (the proportion of biomass between the wings (assuming wing-spread estimates) that is retained in the cod-end). During discussions, areal availability was split into two components: depth and latitude (essentially being the proportion of biomass south of the southern survey-area boundary).

Discussions were held on each of the four components for sablefish, with regard to what was thought to be fully selected fish (being about 53 cm long and perhaps 3-6 years old). The objective with regard to each component was to agree a "lower bound", an "upper bound", and a "best guess". By default, the best guess was the mid-point of the bounds. It was noted that data were available to help with some components (e.g. proportion of biomass south of Point Conception) and finalization of the bounds and best guesses were delayed until the data became available.

The final bounds and best guesses for each component were:

	Depth	Latitude	Vertical av.	Vulnerability
Low	0.85	0.82	0.8	0.4
High	0.98	0.88	1.0	1.0
Best	Mid point	0.85	Mid point	0.8

NWFSC slope trawl survey data from 2003-2006 were used to determine the latitude values. Other values were chosen by consensus (in particular, for the bounds, on the basis that everyone was willing to accept that the "true" value was within the specified bounds).

The consequent bounds on *q* and the best guess are: (0.22, 0.86) and 0.56. The best guess was equated to the median of a lognormal distribution and the bounds to 99% of that distribution. This gave a normal prior on log(*q*): mean = -0.58, sd = 0.184.

The normal prior on log(*q*) was subsequently used to provide three qs for model runs with nominal weights of 25%, 50%, and 25%. A random sample of size 10,000 was generated from the normal distribution and the mean of the samples below the 25th percentile (of the normal distribution) was exponentiated to provide the "low *q*". Similarly, the mean of the samples above the 75th percentile was exponentiated to provide the "high *q*". The median of the prior was used in the base model.

The low, base, and high *q*s were: 0.445, 0.560, 0.712.

Round 3 requests

H: Still to be done (see "Round 2 requests")

J: Four runs – building on progressive base case.

J1: Configuration 4 with discard rate data added (Run A) + tighter sd on annual selectivities (part of A1) + "north of Point Conception" times series (A2) + iterative re-weighting (using ratio of arithmetic means).

J2: Single change from J1: *q* = 0.56 (NWFSC slope survey)

J3: Single change from J2: age based selectivity for the NWFSC slope survey (free up as many parameters as possible)

J4: Single change from J3: estimate *q*

Round 3 responses

Request H was held over from Round 2 requests and required that the "too regular" age-length data be investigated (and corrected). However, the STAT concentrated on request J. An explanation for the "too regular" data was provided: at some time in the past, some age-length estimates had been extended across multiple ages to "get the model working", and the actual data had never put back in.

Run J1 still had the "over the top" problem, as did J2, but to a lesser extent. The results of J3 were not encouraging with poor fits to the age and length frequencies for the NWFSC slope survey. The reason for moving to an age based selectivity for NWFSC slope survey was to ensure that the fixed value of *q* was easily interpretable. There had been a concern within the STAR Panel that the length-age selectivities only reached a maximum of 0.8 when represented as selectivity at age integrated over length (which meant that there were no ages at which all of the fish were fully selected at length).

Verbal requests were made for variations on J3 and J4: removing the NWFSC slope survey length frequencies and ensuring that the age selectivity had initial parameters that made it suitably domed. The variation on J3 resulted in a selectivity that hit bounds on 3 or 4 of the parameters and the run was quickly (but perhaps too hastily) dismissed by the STAR Panel and STAT.

The decision was jointly taken to return to Run J2 as a "progressive" base model. This had the problem that NWFSC slope survey selectivity was age-length based and that the interpretation of *q* was problematic because the length-integrated age selectivity only reached a maximum of 0.8. The STAR Panel was concerned that only a "small locus" of age and length combinations were fully selected. Therefore, the decision was made to rescale the prior on *q* by increasing all values by the reciprocal of 0.8 to account for the expected maximum value of the length-integrated age selectivity.

<u>Round 4 requests</u>

K: Four runs, a base run and three sensitivities (using a rescaled *q* prior based on a maximum age selectivity for NWFSC slope survey of 0.8):

K1: Run J2 with the zooplankton index removed and q = 0.7. Iterative re-weighting must then be completed (this to be done for the age and length frequencies as well as biomass time series for the AFSC shelf survey and the NWFSC slope survey). When the re-weighting is complete, check the fit for NWFSC slope survey biomass. If the "over-the-top" problem occurs for NWFSC slope survey biomass, then the lambda should be increased to 5 (or higher – until the over-the-top problem is resolved). The end result is the base run.

K2: K1 with *q* = 0.556

K3: K1 with *q* = 0.890

K4: K1 with SSH removed.

The decision to rescale the prior on *q* was revisited by the STAR Panel before the Round 4 requests were completed (indeed before K1 was completed). An examination of NWFSC slope survey length frequencies, the observed age-length relationship, and the estimated age-length selectivities convinced the Panel that there were a relatively wide range of ages at which lengths around 53 cm were fully selected. It was decided that the original prior on *q* could stand.

<u>Round 5 requests</u>

L: Further consideration of whether the *q*-prior should be re-scaled or not lead to the conclusion that it should not be (as there appears to be a wide range of ages for which a reasonable number of length bins are fully selected). Therefore, the three runs with fixed *q* need to be done at the original values. The iterative re-weighting and the determination of the NWFSC slope survey lambda should not be redone.

Also, discard data from 2005 are available and should be used for 2005 and future years in preference to the 2004 data.

Three further runs are requested:

L1:	K1 with *q* = 0.445 and the 2005 discard data
L2:	K1 with *q* = 0.560 and the 2005 discard data
L3:	K1 with *q* = 0.712 and the 2005 discard data

Round 4 & 5 responses

Run K1 was partially completed. The STAT reported that after iterative re-weighting, the NWFSC slope survey biomass time series had been substantially down-weighted and the "over the top" problem persisted. After an unsuccessful run with an emphasis level (lambda) of 5, the STAT was uncomfortable with further up-weighting NWFSC slope survey biomass by increasing the emphasis level – wondering what was the point of down-weighting it by one means only to then up-weight it by another. The STAR Panel's explanation, that the initial iterative re-weighting was to provide a "starting point", and that the subsequent up-weighting of NWFSC slope survey biomass was the quickest way to solve the "over the top" problem, was not accepted by the STAT.

Nevertheless, the STAR Panel requested that a run be done with extreme emphasis on NWFSC slope survey biomass simply to see if it would solve the "over the top" problem. While the run was executing, options for other runs were discussed. It was decided to make one last attempt to find a base run which was acceptable to both the STAT and the STAR Panel.

The resolution of the "over the top" problem

A possible cause of the "over the top" problem had been identified by the STAR Panel by pursuing the variation on D3 that had failed to run because of "crash" penalties. Dr Haist had got the variation running and the model had responded by changing the estimated growth parameters. Another variation was run with the growth parameters fixed at their D3 estimates. In this variation, the NWFSC slope survey biomass indices were properly fitted (with a gain of 2 likelihood units) and the likelihood components to suffer (by about 7 units) were the age frequencies from the commercial fisheries. The STAT had, under their own volition, tried a run where the commercial age frequencies were down-weighted in the original base model, but the "over the top" problem had persisted.

The STAT suggested that perhaps the age *and* length frequencies needed to be down-weighted for the commercial fisheries. The potential lack of representative sampling suggested that these data may be problematic. It was decided to remove the additional process error from the AFSC shelf survey and NWFSC slope survey biomass indices and to shift the emphasis factors on the commercial data from 1 to 0.1. The emphasis level on NWFSC slope survey biomass was reset to 1. The idea being to down-weight the problematic data rather than up-weight the data that needed to be fitted properly. The previous day we had chosen the up-weighting option because it was not clear which data sets needed to be down-weighted to solve the problem.

The down-weighting variation worked with an emphasis level of 1 on the NWFSC slope survey biomass. The up-weighting option also solved the “over the top” problem and gave very similar results to the down-weighting variation. The down-weighting variation was accepted as a base model and the definition of the three sensitivity runs was modified accordingly.

Final base model description

The final base model was a modification of the original base model (configuration 4). The changes were:

- Discard rates from Pikitch et al. 1988 and the ADCP database were used and values interpolated as necessary.
- Discard rates from 2005 were used in 2005 and later years (previously, 2004 rates were assumed to apply from 2005 onwards).
- The biomass time series for NWFSC slope survey was replaced by the “north of Point Conception” time series.
- The zooplankton time series was excluded.
- The sd for annual deviations on fishing selectivities was reduced from 1.0 to 0.35.
- The NWFSC slope survey q was fixed at 0.56 (the median of the informative prior).
- Iterative re-weighting was applied to the age and length frequency data sets after which the emphasis levels on the commercial fishery age and length frequencies were set to 0.1 (rather than 1).

It is interesting to note that the original base model gave very similar results to the revised base model. This is coincidental. The original base model had three serious problems: an assumed NWFSC slope survey q=1 (without an adequate basis for the assumption); an “over the top” fit to NWFSC slope survey biomass indices; and inappropriate biomass indices for NWFSC slope survey.

Comments on the technical merits and/or deficiencies of the assessment

The STAT had made significant efforts to improve the assessment for 2007 by simplifying some model assumptions and being more discriminating in the use of data. While this effort was endorsed by the STAR Panel, the Panel is concerned some of the data still included in the model, particularly the length data, add to model complexity without improving the assessment.

The revised assessment is much improved from a technical basis. Its original merits remain but so do its *general* deficiencies. The STAR Panel still retains some unease with regard to the assessment. While we can safely state that it is the best available assessment and we believe it is sufficiently robust to inform management, there is some chance that different results would be obtained if the general deficiencies of the assessment were rectified.

Merits:

- Efforts were made to simply the model and to apply greater discrimination in the use of some data sets.
- SS2 was used and as such brings the advantages of a standard and well tested package.
- Environmental variables were used as recruitment indices which is technically superior to the previous approach (where they modified the stock recruitment relationship).

Deficiencies:

- The complexity of the model is not justified given the likely information content of the available data.
- The use of combined age and length selectivities makes the interpretation of model results extremely difficult. While the concept is not too difficult, the effect that the use of such a complex parameterization has on model results is very difficult to understand. The parameterization also appears unnecessary given that growth morphs are not being used (and so the complexity is imposed simply to fit problematic length data that should probably not be used in any case.)
- Many of the data sets have not been scrutinized and analyzed nearly enough to justify their inclusion in base model runs.
- The age, length, and length-at-age data are used inappropriately. It may not be uncommon to use “all of the data” in this way, but it is technically incorrect. In the case of sablefish it is also unwise. There is almost no genuine information on recruitment (or biomass) in the length data which is not already contained in the age data.
- It was apparent that the STAT had used ad-hoc methods, at unspecified times in the past, to get the model “working.” This had included fixing selectivity parameters and smoothing length-at age input data. Due to an oversight by the STAT, the temporary data were still in the input files in the final runs. Many of the selectivity parameters were also still fixed in the final runs, and there was no documentation for the choice of the fixed values.
- The link between the environmental indices and recruitment remains to be validated (although current results are encouraging).
- A detailed analysis of residual patterns appears not to have been undertaken in recent assessments. E.g., an investigation of sex ratios and whether the patterns are adequately explained by the current model.

Explanation of areas of disagreement regarding STAR Panel recommendations

There were no important areas of disagreement between members of the STAR Panel.

There were two main areas of disagreement between the STAR Panel and the STAT.

The first issue concerned the use of sea surface height (SSH) in the base model. The STAR Panel recommended that it only be used in a sensitivity run. However, the STAT decided to keep it in the base model. As there is very little difference in stock status and projections whether SSH is included or not the dispute is somewhat academic (at least for this assessment). However, the STAR Panel maintain that SSH should not be used as an index of recruitment until a full cross validation study is undertaken and the apparent link between SSH and sablefish recruitment is shown to be robust.

The second area of disagreement was about the *process* used to derive the prior on *q*. In particular, STAT was concerned about the use of the expert opinion from the STAR panel, the GMT and GAP advisors, and the STAT to derive the prior. The STAR Panel certainly agrees that this was not the ideal group nor setting for this task, and that it would be desirable to redo the exercise more comprehensively with a selected group of participants with greater knowledge about fish-trawl interactions and sablefish behavior. There are also data available on fish distribution by depth which could be used to help in determining the depth component of areal availability (the data on fish distribution south of Point Conception was used in the original exercise, but data on depth was not readily available).

Unresolved problems and major uncertainties

As described earlier, the general technical deficiencies of the assessment remain and as such are an unresolved problem. However, as noted, we believe that the current assessment results are probably relatively robust to the technical deficiencies. This is because the assessment is driven by the prior on *q*. We have little confidence that the base model uses the “true” value of *q*, but we are much more confident that the value of *q* is within the range of the prior.

Major uncertainties:

- The value of q remains very uncertain.
- The low-*q* and high-*q* sensitivity runs are only indicative of potential biases in the base model; they do not span the full range of uncertainty.
- There is uncertainty associated with other fixed and estimated parameters including natural mortality and steepness. The implication of errors in these parameters was not explored during the meeting.

Issues of concern raised by GMT and GAP representatives during the meeting

There were no concerns raised by the GMT or GAP representatives that were not addressed elsewhere by the STAT.

Recommendations for future research and data collection

The sablefish assessment needs a full review (this is not possible during a STAR Panel meeting). Additional resources are required to do this. Personnel with specialist experience and skills should critically review each data source. Model complexity should be simplified to be compatible with the expected information content of the data. The starting point should probably be an age-only model with growth estimated outside the model.

Age data, in general, and especially for sablefish, intrinsically contains more information on recruitment (and biomass) than length data. Of course, if ageing methods are unreliable, then age frequencies will be also. The existing age frequencies (and model fits) should be critically examined to see if cohorts (at relatively young ages) are being tracked reliably. If they are not, then ageing methods should perhaps be reviewed (and consideration given to how representative the age samples are likely to be). If cohorts do track reliably, then priority should be given to ageing any remaining samples.

The exercise for deriving the prior on *q* should be redone. All potentially relevant data sources should be made available to a selected group of participants with appropriate skills and experience. Ideally, priors would be formed for all of the trawl surveys used in the assessment. The sablefish q-priors could be derived at a more general workshop covering several species.

The use of environmental variables as recruitment indices is currently fashionable and results do look encouraging. However, the priority for this work is to conduct a full cross validation study on the existing candidates rather than to further refine the candidate environmental indices.

Continuation of trawl time series is essential for future stock assessments. The NWFSC slope survey has been surveying the whole of the Conception stratum in recent years and this should probably continue. If the full survey results are used to construct a time series then the Conception stratum must be subdivided at Point Conception. A consistent time series, using the full area, could be constructed using a number of methods including a GLM or extrapolation using the ratio of average catch rates north and south of Point Conception. A GLM is probably preferable, especially if there are significant vessel effects.

Continued sampling of the commercial fishery is necessary and priority should be given to obtaining *representative* samples (good spatial and temporal coverage for the main fleets).