Southern Black Rockfish STAR Panel Report

National Marine Fisheries Service
Alaska Fisheries Science Center
7600 Sand Point Way NE
Seattle, Washington 98115

October 1-5, 2007

Reviewers

Martin Dorn, Alaska Fisheries Science Center, Panel Chair
Stephen Ralston, Southwest Fisheries Science Center
Andre Punt, University of Washington
Thomas Helser, Northwest Fisheries Science Center
Owen Hamel, Northwest Fisheries Science Center
Tom Jagielo, Washington Department of Fish and Wildlife
Patrick Cordue, Center for Independent Experts

Advisors

Peter Leipzig, Groundfish Advisory Subpanel (GAP) Representative
Kelly Ames, Groundfish Management Team (GMT) Representative

Stock Assessment Team

David Sampson, Oregon State University

Overview

The southern black rockfish assessment was initially reviewed by a STAR panel in May, 2007. The draft assessment model was an innovative attempt to develop a more spatially-explicit approach to assessing black rockfish. Unfortunately model behavior was unstable and it was not possible to resolve the difficulties during the STAR panel meeting.

For this STAR panel review, the STAT retrenched and proposed a preliminary base model similar in structure to the previous assessment, which assumed a unit stock in the waters off Oregon and California. The STAT reconstructed all input datasets and developed a time series of historical catch estimates extending to back to 1915. The assessment also used two new indices of abundance: tagging estimates of black rockfish abundance off Newport, Oregon, and the juvenile rockfish pre-recruit index. However, both of these indices begin after 2000 and are not yet informative about trends in stock status. During the initial presentation of the assessment, the STAR panel learned that the tag abundance estimates had been revised and that STAT's preliminary base model now included those revised estimates. This change had a minor impact on assessment results.

A primary focus of the review was how length at age was modeled in the assessment. The Panel was concerned about the lack of fit to pre-recruit index when there was no apparent additional information to inform the model about recent recruitment patterns. SS2 uses linear extrapolation to extend the growth curve below the Amin growth parameter, which was set at age 3 in the preliminary base model. The Panel surmised that this could produce an unrealistic growth curve for the younger fish that was interfering with the fit to the pre-recruit index. To address this problem, Amin was lowered to age 1 and an additional size bin was added to the lower end of size composition data. Alternative parameterizations of the CVs of length-at-age were also evaluated, and a new value was adopted (0.07 for all ages) based on improved model fits. The modified treatment of growth did somewhat improve the fit to pre-recruit index, but substantial lack of fit remained. The STAR panel ultimately concluded that the lack of fit could be explained by 1) the low level of recruitment variability, 2) the low emphasis given to pre-recruit index during iterative re-weighting of input variances, 3) an inconsistency between the index and size composition data, particular the California RecFIN data, which did not show the presence of a mode (or a shoulder) of small fish consistent with high pre-recruit index values.

The Panel also attempted to understand the tradeoff between the fits to various data sets used in the assessment. Most of the length data fit best at high estimated stock size, while age data fit best at low stock size. Length-at-age data, which ordinarily should not be informative about stock size, was also influential, and fit best at high stock size. Attempts were made to reduce the influence of the length-at-age data by combining all the data in a single year, and by iterative re-weighting. While this reduced the influence of the length-at-age data, the fit to these data showed a pronounced bias for the older fish. The STAT and the STAR panel agreed a good fit to length-at-age data was important to preserve, and so the final model did not use iterative re-weighting for the length-at-age data. The final model is an attempt to balance opposing datasets that imply differing views of stock status, which increases the overall uncertainty of assessment results.

Finally, the STAR Panel recommended a number of relatively minor changes to "tidy up" the model. These changes included a revised ageing error matrix based on double reads from the set of consistent age readers, the addition of a catchability break in 2005 in the Oregon CPUE index to model the most recent change in the bag limit, and the use of a new value for the recruitment variability parameter (σ_R) based on a single iteration (i.e., it was not iterated to convergence). Further, the start year for estimating recruitment deviations was changed from 1975 to 1970. None of these changes had a substantial impact on assessment results.

Analyses requested by the STAR Panel

Round 1 requests

New runs are all variations on the existing base model with the revised tag estimates. For each new run present any significant changes to fits or estimates.

A: Present diagnostics for the GLM analyses (e.g., residuals vs predicted values, qq-plots, or other diagnostics that illustrate whether the assumptions of the models are satisfied).

Reason: To evaluate whether the GLM model assumptions are satisfied or not.

Response: Diagnostic plots were presented for the positive-catch sub-models for Oregon and California. The Oregon model used a gamma distribution and the standardized residuals were not symmetrically distributed and not centered on zero. The California model used a normal distribution and the residuals were reasonably symmetric but showed some departures from the assumption of normality.

Discussion/conclusion: The Panel had difficulty interpreting the diagnostics presented for the gamma model. It was concluded that alternative diagnostics were needed to draw valid conclusions about whether the assumptions were met or not. It was suggested that a general recommendation be made that a standard set of easily interpretable diagnostics be routinely calculated (and presented) for all GLM analyses. This will require that appropriate diagnostics are identified for each type of GLM analysis (e.g., gamma, binomial, or lognormal) and that suitable tools are made available to STATs.

B: Present a year-by-year comparison of the length frequencies for aged and un-aged fish (with sample sizes). Scaled length frequencies should be used in both cases if feasible.

Reason: To evaluate whether the aged fish were selected at random or not.

Response: Comparisons were presented for the ORBS dataset using un-scaled length frequencies. The aged and un-aged fish had very similar length frequencies in most years.

Discussion/conclusion: It was not clear how best to obtain scaled length frequencies for this comparison. However, there is no evidence that the aged fish are unrepresentative of the fish that were sampled for length.

C: Tabulate selectivity parameters indicating which were freely estimated, which were fixed and why.

Reason: To clarify exactly what was done.

Response: A table was presented showing the required detail and also whether parameters were estimated at a bound or not. All parameters of the mainly double normal selectivities had been estimated freely. Some double normal parameters were on the upper bounds (8/36), but half of these were for the selection in the lowest length bin (i.e., no selection of fish in the smallest bin).

Discussion/conclusion: There was discussion about why the parameters on bounds had not caused problems for inversion of the Hessian. There were mixed views on whether it should cause problems or not. It was suggested that the parameters on the bounds should be fixed at those values, but it was agreed that if this is done, it should not be until a final base model has been chosen.

D: For the set of otoliths that were read by three of the standard readers, plot each individual reading vs average age.

Reason: To evaluate how much spread in age readings there is amongst the three readers.

Response: It was reported that some of the duplicate readings used to calculate the ageing-error matrix had been from non-standard readers (these should have been eliminated). Duplicate readings were actually only available for two standard readers. These results were presented on the plot as requested. They showed general agreement but also some otoliths which had been assigned very different ages by the two readers (e.g., 5 years vs 10 years).

Discussion/conclusion: A minor revision to the ageing-error matrix was noted. The between-reader variability was not considered to be unusual.

E: New run: separate time series (q) for the 2005 and 2006 points in the Oregon recreational CPUE time series.

Reason: To disconnect the 2005 and 2006 indices from the time series because of much more restrictive bag-limit regulations in 2005 and 2006. The points were retained with a different q rather than being deleted so that future updated assessments would be able to include the 2005 and 2006 indices as well as future indices.

Response: As expected this minor change was of no consequence.

F: Explore sensitivity to CVs of length-at-age (or report previous results).

Reason: To check for potential sensitivity to CVs of length-at-age.

Response: Three parameters (cv-young, cv-old-female, cv-old-male) were varied to produce an array of alternative runs. The best fit was obtained when all parameters were equal to 0.07 (which gave an improvement of 30 likelihood units over the provisional base model). The assessment results were not sensitive to the parameters.

Discussion/conclusion: The STAT recommended moving to a model with the CVs of length-at-age fixed at 0.07. There were no objections.

G: New run: block selectivity in the same way as CPUE qs.

Reason: The CPUE time series were split because of regulation changes which could have impacted on selectivity as well as catchability.

Response: This change was made cumulative with the split in the Oregon CPUE recreational time series. Four parameters were estimated for each block's selectivity. There was an improvement of 50 likelihood units for 12 parameters. There was not much difference in the estimated selectivities in Oregon, but there were some differences in California.

Discussion/conclusion: There was a general feeling from meeting participants that the changed selection patterns were contrary to what would have been expected given regulation changes and changes in fishing patterns. Blocking made little difference to the assessment results and it was agreed not to use it in any base models.

H: Evaluate using mirrored selectivities for data sets sampling the same recreational fisheries (e.g., are the splits justified in terms of a decrease in likelihood units?)

Reason: To determine if the splits are justified in terms of an improved fit.

Response: The mirrored run had a much poorer fit to the data overall (a drop of 200 likelihood units) with the main degradation in the recreational length frequencies as expected.

Discussion/conclusion: It was concluded that the splits were justified.

I: Explore why the pre-recruit time series is so poorly fitted (e.g., try a range of larger values for σ_R).

Reason: It was not clear why the pre-recruit time series could not be fitted exactly (as it had been in the blue rockfish assessment).

Response: The STAT described a process where he had cumulatively made changes to try to obtain a good fit to the pre-recruit time series. At the end of the process the 2005 and 2006 length frequencies were removed from the likelihood, the pre-recruit indices had been converted to an "arithmetic scale" (incorrectly as they were already in arithmetic space) and a larger $\sigma_R = 0.75$ was used. After all of these changes, the fit was improved but was still missing the 2004 index.

Discussion/conclusion: It was suggested that there was an interplay with the relatively large CV of length-at-age on young ages so that the pre-recruit series was competing with length frequencies. This was confirmed and a suggestion was made to specify zero selection at age 1 (and perhaps 2). There was also a suggestion to use option 33 (post density dependent) rather than 32 (pre-density dependent) for fitting the pre-recruit time series. It was concluded that the current growth curve had unrealistic sizes at ages 0, 1, and 2 years (being too large). This was all to do with SS2 using the lower edge of the lowest growth bin in the growth curve. Various suggestions were made on how to make the growth curve more realistic (without too much effort). The proposal to specify zero selection at young ages was warned against as there were no age data from California (perhaps young fish were being caught).

J: Explore what is driving the estimates of the two strong year classes.

Reason: There is little visual evidence of strong cohorts in the data.

Response: This was explored at various times during the meeting by the STAT but was accorded lower priority than other requests. Some results were presented where the strong recruitment estimates had been constrained by bounds. This showed degradation in the fit to some length frequencies.

Round 2 requests

K: Construct a new candidate base model and profile across R_0.

Make the following changes to the preliminary base model:

- Ensure that the growth curve has realistic mean length-at-age for 1 and 2 year old fish.
- Use the new ageing-error slope
- CVs of length at age = 0.07

Reason: A new candidate base model was needed to progress the assessment. The important change was to incorporate realistic mean length-at-age for the young fish. There was concern that inappropriate sizes for the young fish could distort the assessment results (through poor recruitment estimates). See discussion for Request I.

Response: An extra length bin was inserted with zero observations for each length frequency (thus reducing the lowest length in the length bins). Also, the Amin parameter from reduced from 3 years to 1 year. The pre-recruit series was still not fitted.

Discussion/conclusion: It was agreed that the new growth curve was realistic at young ages. There was further discussion about options 32 and 33 and it was agreed to try option 33 (see Request M). It was concluded that the length frequencies must be competing with the pre-recruit time series (even given the new growth curve).

Round 3 requests

L: Explore candidates for a new base model.

Use the model with cumulative changes from request K and request E:

- Choose a start year for estimating recruitment deviations by considering the standard deviations of recruitment deviations.
- Do runs with three alternative starting values of σ_R: 1.0, 0.8, and 0.5, and compare with output values.

Reason: To finalize σ_R and the start year for estimating recruitment deviations in the base model.

Response: The plot of the standard deviation of recruitment deviations suggested that 1970 was an appropriate start year (although 1975 wasn't too bad). All output values of σ_R were lower than the input values (1.0 gave 0.47; 0.5 gave 0.37).

Discussion/conclusion: It was agreed to start estimating recruitment deviations from 1970. In the spirit of letting the "data speak" (somewhat), σ_R was fixed at 0.5 (being the approximate output from an input of 1.0).

Round 4 requests

M: Explore candidates for a new base model.

Use the model with changes from request L:

- Use option 33 for fitting the pre-recruit time series.
- Fit only one year of the mean length-at-age data (choose the year with the largest sample).

Reason: To finalize the option for fitting the pre-recruit time series (pre- or post- density dependent effects) and the mean length-at-age data. There was a concern that the three years of mean length-at-age data perhaps showed a trend which the model was attempting to fit (and hence giving the data undue influence on R_0).

Response: Use of option 33 gave a slightly worse fit to the pre-recruit time series and no effect on results otherwise. Data from 2004 were slightly more numerous so it was chosen as the single year for mean length-at-age data. The new run had somewhat lower estimates of depletion and MSY and had an improved fit to the length and age data. There was little reduction in the mean length-at-age likelihood component (despite there being only one year of data). This was because there was a poor fit to the 2004 mean length-at-age.

Discussion/conclusion: Option 32 was agreed upon for base models. There was concern about the poor fit to the mean length-at-age data and it was agreed to explore other options for using the data (see Request N).

Round 5 requests

N: Refine base model.

Use the model with changes from request L. Explore options for the use of the mean length-at-age data with the objective of removing the strong contrast in the likelihood component with respect to R_0.

Reason: Continued concern that, in reality, mean length-at-age data are not informative about abundance.

Response: Two options were tried: 2003 only, and all years combined and input in 2004 ("composite" run). Tuning was done to three iterations which resulted in down-weighting of the mean length-at-age data. Profiles over R_0 were shown and the mean length-at-age data showed a much reduced contrast compared to earlier runs.

Discussion/conclusion: It was noted by the STAT that the reduced contrast from the mean length-at-age data was partly due to the tuning process and subsequent down-weighting. The same tensions were present between the data sets as in earlier runs. The composite mean length-at-age run was preferred by the STAT and the Panel agreed (it had "more data").

O: Explore dimensions of uncertainty.

With the base model from Request N do high and low runs with respect to M and catch history, first individually and then in combination. The low and high vectors for M are respectively, (0.12, 0.18) and (0.19, 0.28) – the same values used for northern black rockfish.

Reason: The estimates of depletion and MSY were likely to be most sensitive to these dimensions of uncertainty.

Response: The requested runs were presented, with the catch histories only including low and high elements for trawl. The low and high runs for M showed strong contrast in MSY and depletion. When combined with catch the contrast increased somewhat.

Discussion/conclusion: There was some discomfort with the results, with the low and high combinations being considered too extreme to represent the "ideal" of 25% "probability" each (with 50% probability on the base model). The CV on the MSY estimate for the base model was requested and found to be approximately 10%. It was noted that this greatly under-estimated the "true" uncertainty (model uncertainty being the key element). It was suggested that further investigation of uncertainty was warranted.

Round 6 requests

P: Explore dimensions of uncertainty.

With the base model from Request N do high and low runs with respect to M and catch history in combination. Use a low and high M for young fish of 0.14 and 0.18 respectively (with the base model offset for old females). Also include the low and high trawl-catch histories.

Reason: An attempt to find less extreme runs meeting the ideal of 25% probability each for the low and high runs.

Response: Before completing the request the STAT requested that the choice of base model be revisited. His concern was that the fit to the mean length-at-age data was very poor; so much so that the estimated growth in the model was inconsistent with the "reality" being represented by the mean length-at-age data (which he considered was much more reliable for estimating growth than the other data in the model). He proposed returning to an earlier model than that from Request N, where the mean length-at-age data were somewhat better fitted (because they had not been down-weighted due to tuning).

Discussion/conclusion: There was general agreement that there was a problem. There was an inconsistency between the mean length-at-age data and other data sets in the model. In particular, the conflict was with the Oregon age frequency from which the mean length-at-age data were derived. Balancing the STAT's concern that inappropriate growth was being estimated in the model was the concern that if too much weight was given to the mean length-at-age data then it would have an undue influence on abundance estimates. It was agreed that an earlier run would be used but that a sensitivity would be done to the weight on the mean length-at-age data (i.e., lambda = 0.1).

Response: The low and high runs from the new base model provided a less extreme range for depletion and MSY than those from Request O. The sensitivity run with lambda = 0.1 on the mean length-at-age data gave results very similar to the low run.

Discussion/conclusion: The Panel and the STAT agreed that the low and high runs conformed to the 25% probability guidelines. They also agreed that some cautionary words should be used to note the sensitivity of results to the weight placed on data which were primarily meant to inform growth.

Final base model description

The final base model was a modification of the preliminary base model. On the whole, the changes were relatively minor and include:

- The growth model in SS2 was re-specified by lowering Amin to age 1 and an additional size bin was added to the lower end of size composition data.
- The CV of length at age was set to a constant value of 0.07 for all ages.
- Length-at-age data was entered for a single year to reduce its impact on the estimate of stock size.
- The Oregon ORBS CPUE index was split with a new catchability period in 2005 to account for the most recent change in the bag limit.
- The ageing error matrix was revised based on double reads from the set of consistent age readers.
- A new value for the recruitment variability parameter (σ_R) was adopted based on a single iteration from a starting value of 1.0 (i.e., it was not iterated to convergence).
- The start year for estimating recruitment deviations was changed from 1975 to 1970.

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

<u>Technical Merits</u>

- The assessment generally achieved a good balance between model complexity and data availability.
- Inclusion of tagging estimates of abundance is a positive step.
- SS2 was used effectively to model population dynamics, growth, and size-specific fishery impacts. SS2 brings the advantages of a standard and well tested package.
- Age data were carefully evaluated for consistency before they were added to the model. Questionable data were excluded.
- Substantial improvements were made to the historical catch estimates. In particular, the parallel development of low, medium and high estimates was an important advance.

<u>Technical Deficiencies</u>

- More appropriate ways of using both age and length data are available, i.e., conditional age-at-length compositions.

Areas of disagreement regarding STAR Panel recommendations

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

Unresolved problems and major uncertainties

- There was a substantial change in estimated stock size between this assessment and the previous assessment. This in itself raises concerns about the temporal stability of assessment results.
- The assessment area is based on management boundaries and not on population structure.
- Historical catches of black rockfish are highly uncertain.
- Various datasets used in the assessment do not provide a consistent indication of stock status. Estimates of current status are a result of balance between conflicting datasets.
- Natural mortality is fixed to same values used in the northern black rockfish assessment. In reality, natural mortality is highly uncertain and cannot be reliably estimated.
- The assumed value of stock-recruit steepness was based on Dorn's meta-analysis of steepness and represents average for all West Coast rockfish. The assessment itself provides little indication of the appropriate value of steepness for black rockfish. Consequently, how the stock will respond to the Council's harvest policy for rockfish is not well known.
- The assessment only considers black rockfish status at the stock level. Are there local areas that have been fished more intensely and thus become more depleted than the stock as a whole? This question is not addressed in the assessment.

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

There were no issues of concern raised by GMT and GAP representatives during the meeting.

Recommendations for future research and data collection

- Additional work is needed to develop a quantitative prior for tagging catchability. Tagging catchability should be based on analysis of potential black rockfish habitat and the relative abundance of black rockfish throughout the geographic range of the assessment (see Appendix IV to the 2005 cowcod assessment). Continuation and/or expansion of tagging programs should consider the scope of project the relative to the area being assessed. If the area covered by the project is small relative to assessed area, the potential to provide useful information for stock assessment is limited. Development of priors for tag catchability should consider uncertainty as well as point estimates.
- Development of a fishery independent time series using fixed sites and volunteer fishers properly supervised using standard protocols. The CPFV dataset consisting of reef-specific CPUE data has been repeatedly identified as most valuable index for monitoring stock trends of nearshore species.
- The STAT excluded a large amount of ageing data because of inconsistencies that made it unsuitable for use in the assessment model. This raises concerns about

age reading protocols. Age reader comparisons, both between readers within the same agency and between readers from different agencies, should be a routine part of age reading procedures.

- This assessment was limited by inadequate biological sampling of California component of the recreational and commercial fishery for black rockfish. Recreational fishery length data could not be expanded to landings because strata with large landings were not sufficiently sampled. Age data were unavailable for California, which made it impossible to compare geographic differences in growth. There have been positive steps towards sustainable management of nearshore species off California at the policy level, but the lack of investment in long-term sampling programs for biological data may make it difficult to achieve policy objectives.
- For stocks whose primary assessment index is derived from recreational fishery CPUE, greater consideration should be given to the potential impact of management changes on the ability to assess the stock. Management tools such as bag limit and season closures may have different impacts on CPUE trend data. Each management change, e.g., a bag limit change, potentially reduces the value of fishery-dependent data.