Center for Independent Experts (CIE) Report on the 2014 South East Data, Assessment and Review (SEDAR 33) Gulf of Mexico Gag and Greater Amberjack Assessment Review, 24-27th February, Miami, Florida

Michael J Armstrong[1]
Centre for Fisheries, Environment & Aquaculture Science (Cefas)
Pakefield Road
Lowestoft
Suffolk NR33 0HT
United Kingdom
mike.armstrong@cefas.co.uk

[1] Representing the Center for Independent Experts.

1. Executive Summary

The 2014 South East Data, Assessment and Review (SEDAR 33) Workshop to review the assessment of the status of Gulf of Mexico Gag and Greater Amberjack took place on 24-27th February 2014 in Miami, Florida. The material presented was derived from a separate Data Workshop and an Assessment Workshop, some of which was updated with new information following the Assessment Workshop. The primary assessment model for each stock was Stock Synthesis, which is a length and age structured statistical model using fleet-specific commercial and recreational landings, discards, length and age compositions, and a range of fishery-dependent and fishery-independent abundance indices, and fits fishery selection curves and retention ogives by fleet. The previous SEDAR assessments used a different statistical model for Gag grouper (CASAL) and a dynamic surplus production model (ASPIC) for Greater Amberjack. The Assessment Workshop was required to provide continuity assessments showing the outcomes of using the previous assessment models with updated data series. For Gag grouper, this was achieved by configuring Stock Synthesis to replicate, as far as possible, the structure and parameterization of the CASAL model.

For both species, a comprehensive review and analysis of data sets had been carried out to provide inputs to the proposed Stock Synthesis models and continuity models, and the assessment team had carried out a very wide range of sensitivity analysis runs prior to the Review meeting, which helped enormously in evaluating the quality of the assessment, projections and status evaluation.

For Gulf of Mexico Gag, the Review Panel accepted the Stock Synthesis model as a basis for providing advice although recommended the use of a fixed stock-recruit steepness value of 0.85, with the much higher Stock Synthesis estimate (as presented by the assessment team) and a lower value of 0.7 to indicate sensitivity of advice to uncertainty in this parameter. The perception of whether the stock of this protogynous hermaphrodite species is overfished is very sensitive to how SSB is calculated. The female-only SSB model indicates that the stock is no longer overfished in relation to any of the proposed reference points. However, the SSB-combined (male plus female) model indicates that the stock is overfished in relation to SSBSPR30%, but the SSB is marginally above SSBMSY. The stock does not appear to be undergoing overfishing. Recent management measures aimed at reducing the exploitation rate appear to have successfully lowered recent F values to below Fmsy and also Fspr30 for the agreed base model. The reduced IFQs in particular have caused a large increase in discard rate across the size range, and the benefits of the measure in part depend on survival rate of discards. However the estimated reduction in F, including the assumed discard mortality, is very large and unprecedented for the time series.

For Gulf of Mexico Greater Amberjack, the Review Panel concluded that Stock Synthesis was a more appropriate modeling framework than the previously used ASPIC model for this stock, as it can estimate changes in selectivity and retention ogives which affect the composition of catches in a way that ASPIC cannot interpret. However, poor Stock Synthesis fits for many of the fishery composition data series, and unstable solutions, indicated model or data problems that could not be resolved during the Review Meeting. The Review Panel members were all of the opinion that Stock Synthesis remained an appropriate assessment approach for this stock, but the optimal configuration of the model had not yet been found.

As no base case assessment model could be identified, projections and stock status could not be evaluated. The Reviewers' identified issues that should be addressed before the assessment model could be accepted as properly configured and consistent with standard practices, and considered that these could feasibly be addressed after the Review Meeting to provide advice this year.

2. Background

South East Data, Assessment, and Review (SEDAR) is a joint process for conducting stock assessments, and peer-reviewing their outcomes, for stocks of interest to the South Atlantic, Gulf of Mexico and Caribbean Fishery Management Councils, NOAA Fisheries, SEFC, SERO and the Atlantic and Gulf States Marine Fisheries Commissions. SEDAR is organized around separate data, assessment and review workshops. The previous assessment of Gag was conducted at the SEDAR-10 Update (2009), and for Greater Amberjack the previous assessment was the SEDAR-9 Update (2010). Input data for the SEDAR-33 assessment were compiled during the Data Workshop (DW), and population models were developed during the subsequent Assessment Workshop (AW).

3. Description of reviewer's role in the review activities

The SEDAR 33 Review Workshop (RW) took place at the DoubleTree by Hilton hotel in Miami, Florida, from 9:00am Monday 24 February to 12:00am Thursday 28 February, 2014. The assessment results and background were clearly presented by the experts at the meeting. The Review Panel requested a number of additional Stock Synthesis model runs for each stock to explore the sensitivity to decisions made by the Assessment Workshop, and to explore the outcomes of changes to model settings suggested by the Panel. The sensitivity analyses were done very quickly and led to fruitful discussion that helped to clarify a number of important issues, including the robustness of the proposed base-case models. The provisional agenda for the meeting is given in Annex 3 of Appendix 2.

The Review Panel itself comprised the Chair, three reviewers appointed by the CIE and a Science and Statistical Committee representative. The assessment results were presented by four technical experts who were involved in the AW. The RW was also attended by the SEDAR coordinator and a number of SEFSC, SSC and Council staff members. All documentation, including background documentation provided to earlier DW and AW meetings, was provided to the Review Panel in advance of the review workshop, and was comprehensive for the job in hand.

The Review Panel provided a Summary Report. The following report presents my personal evaluation of the review process together with more extended observations on the data and assessment models. The Panel achieved a consensus view on all findings documented in their report. However my more extended observations are not necessarily shared with the other Panel members. I accept all responsibility for any errors in my report due to misinterpretations of the data or analyses.

4. Summary of findings by Term of Reference

4.1 Gulf of Mexico Gag

My evaluation of the assessment and review process are given below by Term of Reference for the review process. This includes many outcomes documented in the draft of the Review Panel report, which was compiled collaboratively by the Panel, and is expanded where necessary with my own more detailed opinions and recommendations. In places I have replicated the Review Panel text where it clearly represents my own views.

ToR. 1. Evaluate the data used in the assessment, addressing the following:

a) Are data decisions made by the Assessment Workshop sound and robust?

With some exceptions, the decisions made by the Data and Assessment Workshops were sound and robust, and the data used in the assessment are adequate and appropriate for that purpose. The Stock Synthesis modeling framework was chosen to allow incorporation of a range of fishery and abundance index data by length and age, in a way that can account for changes in fishery data caused by management actions affecting selectivity and retention (discarding). The Data Workshop conducted an extensive compilation and evaluation of all available fishery dependent and independent data sets, and reviewed methods for determining input values for natural mortality and discard mortality, and for estimating probability of maturity and transition from female to male. Decisions were made on how to make imputations where data were missing or inadequate; approaches for extrapolating data series back in time; and inferring historical Gag catches from mixed-species catch records. Fishery-dependent abundance index data were filtered using species guild criteria in an attempt to remove trips in areas beyond Gag habitat. All fishery-dependent and independent abundance data were analyzed using delta – lognormal modeling to remove the effect of a number of factors other than abundance affecting interannual changes in observations such as fishery catch rates or video camera counts. Overall, the raw data were subject to extensive processing to provide assessment inputs.

All decisions were well explained in the workshop reports and presentations to the Review Panel. In most cases the decisions were sound and robust, and where appropriate their effect on the assessment was adequately explored in sensitivity analyses. However, in my opinion, the Assessment Workshop made two inappropriate decisions:

- Firstly, recreational catches were treated as exact, when in fact they are estimated from probability-based surveys which are designed to minimize biases as far as possible and to provide robust estimates of precision. This is in contrast to commercial landings figures which are based on census (i.e. very high precision) but may have biases due, for example, to problems with species reporting. Statistical models should be fit taking into account the known precision of the input data (where random measurement error is the main source of uncertainty), whilst biases should be explored through sensitivity analysis.
- Secondly, a decision had been made to use numbers of fish measured or aged, capped at 200, as input effective sample sizes (ESS) for fishery composition data. It is well known that numbers measured or aged is a poor indicator of relative precision due to cluster sampling effects. The use of individual fish numbers resulted in the

composition data for many fleet-year combinations being capped to avoid very high weighting of some years and fleets. Stock Synthesis generates output ESS values based on the fit to each year's data, but extensive capping of input ESS precludes the ability to examine the correlation between input and output ESS. If poorly correlated, but the input ESS correctly reflect the relative precision of the input data, it would suggest that factors other than random sampling error are degrading the model fit to the data. The Review Panel strongly recommended that if the true ESS is not calculated, a proxy should be developed, for example the number of independent primary sampling units (such as numbers of trips sampled). During the Review Meeting, it was commented that sampling in the earlier years tended to involve fewer trips and more fish measured per trip than in later years.

Additional assessment model runs were requested at the Review Meeting to treat the recreational catch estimates as subject to survey error rather than being exact, and using numbers of trips sampled for length or age as proxies for effective sample sizes, without any capping. These did not alter the perception of recent stock trends and status, and whilst the Panel agreed that this remained a desirable approach, further work would be needed to identify the most suitable model inputs on effective sample size and recreational survey precision.

b) Are data uncertainties acknowledged, reported, and within normal or expected levels?

Data uncertainties were explored and reported, although it would have been more helpful to see a clearer summary of the relative quality of the different data sets. Where they could be reliably quantified, uncertainties appeared to be within normal or expected levels given the design of data collection schemes and the amount of sampling that has taken place. As mentioned under ToR1(a), the raw data were subject to extensive processing to provide assessment inputs. Many of the approaches adopted, such as imputations for missing data for years, fleets or areas, corrections for mixed-species reporting, filtering of relative abundance data etc. can themselves lead to variable bias over years. The Data and Assessment Workshops acknowledged the major uncertainties and described how they dealt with these. However, it was difficult as a reviewer to clearly understand the relative quality of the different data sets, either from the workshop reports or during the review process. In some cases, for example the recreational fisheries (which take a large fraction of the catches), the data are collected using statistically-sound sampling designs to reduce bias and allow more accurate estimates of precision, and the quality of the data are easier to evaluate. In such cases, uncertainties appeared to be within normal or expected levels given the design of data collection schemes and the amount of sampling that has taken place. In other cases, for example commercial fishery sampling at sea or on shore, filtering of relative abundance data from commercial fleets, and imputation or hind casting of missing data, it is not so clear what types and magnitude of bias might be present.

In general, a clearer framework for documenting known or potential data quality issues (bias and precision) in relation to design, implementation, sampling achievement and analysis of data over different periods, using suitable quality indicators, would be very helpful for assessment analysts and reviewers. Evaluating data quality through performance in an assessment model is not sufficient in itself if the errors in the data include biases as well as sampling variance. Although the relative abundance series were subject to detailed statistical modeling, I felt that more consideration could have been given to *a-priori* evaluation of whether the data are capable of providing indices directly proportional to abundance. For example, a number of factors such as season were included in the commercial longline index computation (Cass-Calay: AW07), but no discussion was included on potential effects of other factors such as competition and saturation of hooks by competing species (which may have different trends in abundance to Gag), or technological changes that may have altered catchability over time.

c) Are data applied properly within the assessment model?

Based on the workshop reports and presentations at the Review Meeting, the data were properly applied within the Stock Synthesis assessment model.

d) Are input data series reliable and sufficient to support the assessment approach and findings?

The input data series are in general reliable and sufficient to support the assessment approach and findings. In this context I interpret “reliable” as meaning an acceptably low bias and variance, for example fishery sampling that is representative and achieves sufficient numbers of primary sampling units. Reliable survey or fishery cpue series should adequately cover the stock and in principle be capable of being proportional to abundance. Not all years of Gag fishery data have adequate length or age data, particularly in the early part of the time series, and some surveys cover only part of the stock. However, the deficiencies and uncertainties in the data have (in most cases) been explored in detail, and assumptions and decisions made in compiling input parameters and data are clearly presented and their effect on the assessment shown through sensitivity analyses. Good fits to the data and the ability of Stock Synthesis to find a stable solution also indicate that the data are sufficient to support the additional complexities in the model structure around selectivity and retention that are necessitated by the series of changes in minimum landing sizes and IFQs which affect size compositions of retained and discarded fish.

2. Evaluate the methods used to assess the stock, taking into account the available data.

a)-c): Are methods scientifically sound and robust? Are assessment models configured properly and used consistent with standard practices? Are the methods appropriate for the available data?

The Stock Synthesis 3 (Methot 2013) modeling framework is widely used in the USA and elsewhere and is well tested in peer-reviewed assessments. Its strengths include the ability to fit to mixtures of length and age composition data that can include missing years, and to explicitly model the selectivity and retention patterns of component fisheries and any changes in these over time. This is important and appropriate given the nature of the gag data and the major changes in management measures affecting retention (discarding) practices. The approaches adopted to model selectivity (constant over time) whilst allowing changes in retention ogives to account for changes in discarding practices, the effects of which were evident in the catch composition data, were statistically sound and robust.

The assessment team fulfilled their remit to carry out a continuity assessment, not by updating the previous CASAL assessment, but by configuring Stock Synthesis to closely mimic it. In this configuration, Stock Synthesis gave very similar, but not identical, results compared with the previous CASAL assessment. On the basis of this finding, the developments of the Stock Synthesis application to Gag in SEDAR 33 are an improvement over the previous assessment.

The Stock Synthesis model had to be configured to allow for the fact that Gag are protogynous hermaphrodites (female at birth, then a proportion of the population transition into males). This process was elegantly included in the model as a (fixed) logistic function expressing the probability of being a female at each age. The dynamics and drivers of the trigger to change sex are poorly understood at present, and further work is warranted given the importance of this in relation to the choice of SSB metric for management.

An additional non-standard feature to be modeled was the known increased mortality caused by the red tide event in 2005, which is evident as a sharp decline in most abundance indices. The decision to include this mortality as an additional "fleet", which only has positive effort in 2005 and flat selectivity for all ages greater than zero, appeared the most defensible amongst alternative approaches investigated. This was a novel approach to estimating the mortality caused by red tide, and is worthy of publication in the primary literature.

3. Evaluate the assessment findings with respect to the following:

a) Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?

The final SS3 abundance and biomass estimates represent the best fit to a wide variety of input values for landings, discards, length and age compositions, and abundance indices, and the estimates of exploitation rate are derived from the ratios of catch to abundance conditional on assumptions regarding the shape of selection curves. As discussed above, the relative quality of the different inputs are in some cases hard to interpret. Input CVs may represent apparent precision but may be only a component of variability if there are serially correlated biases in data sets that are not accounted for. Based on the jitter analysis, the Gag assessment appears to converge robustly on a solution, but the fits to the different data sets were quite variable. The model was set up to fit catches exactly (relaxing this for recreational catches was explored in an additional sensitivity analysis requested by the Review Panel), and in general the catch composition estimates were fitted quite well. In contrast, there was a tendency for the relative abundance indices to show non-random (serially correlated) residuals, with some common patterns suggesting some conflict between the abundance indices and the catch data.

The commercial and recreational fishery abundance indices were derived as standardised cpue series, and are essentially the same trip catches and their associated effort as for the catch data used in the assessment, albeit modelled at the trip level with extensive filtering of trips. The fleet selectivity patterns estimated in SS3 are also applied in generating expected values for the abundance indices. Hence the catch and abundance index series are not completely independent. The model fitted poorly to the two commercial abundance indices (hand lines and longlines), both of which gave indices persistently below the SS3 estimates in the 1990s and persistently above the SS3 estimates in the 2000s (Fig. 1 below – from Review Panel presentation), indicating biased trends in the indices and/or the SS3 estimates. The model fitted best to the fishery-dependent headboat index, influenced by the relatively low CV for that index. That index partly contributed to the model prediction of a steep increase in biomass in recent years, suggesting the input of recent strong year classes given the domed selection for recreational fisheries peaking below 5 years. This increase is not so evident in the other recreational fishery indices, which have similar domed selection curves.

The fits to the fishery-independent video indices are poor, although the series are short. One of these surveys, the PC-video, covers only part of the along-shore range of the stock. The longer (more continuous) series of 0-gp sea grass survey indices also show a period of negative residuals in the 1990s switching to positive residuals in the 2000s (Fig. 1). Stacking the residual plots suggests a consistent tendency for positive residuals in the 2000s, peaking in the middle of the decade (lower right hand plot in Fig.1). The consistent residual patterns in a wide range of different abundance index series suggests the problem may lie in the fit to the fishery catch and catch composition data, which are heavily weighted in the model.

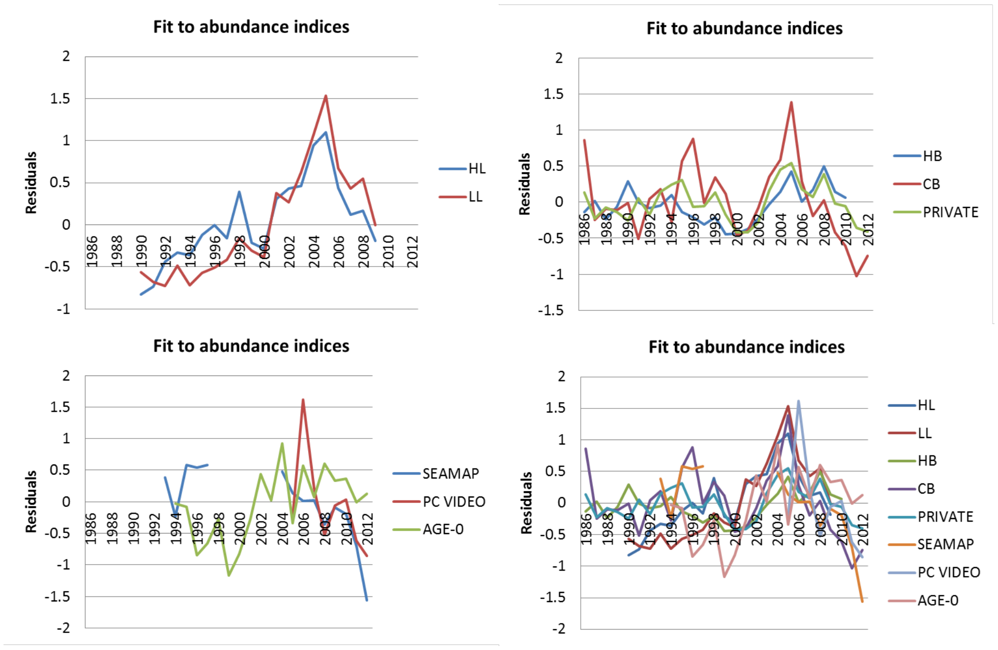

Fig. 1. Residual patterns from base-case SS3 model fit to relative abundance series: commercial handlines (HL) and longlines (LL), recreational headboats (HB), charter boats (CB), private boats (PB) and the fishery independent indices from Seamap, PC-video and age-0 sea grass survey.

The Panel report notes that, if additional iterative reweighting, for example, had been applied, then most or all of the index CVs would have been adjusted upwards, thus giving the abundance indices even less weight in the model. Recent advice in the fisheries literature on model weighting (Francis, 2011[1]) recommends that abundance indices be weighted more heavily than iterative weighting of all data sources would normally provide. A number of SS3 sensitivity runs were carried out to explore the effect of down-weighting the catch composition data and allowing an input CV for recreational catches representing the calculated precision of the estimates. Heavily down-weighting the catch composition data resulted in an overall increase in SSB and

[1] *Canadian Journal of Fisheries and Aquatic Sciences*, 2011, 68(6): 1124-1138

reduction in F across the series, but little change to recruitment. The perception of a recent sharp increase in biomass is not altered.

The Stock Synthesis implementation for Gag estimates numbers at age from age zero, and is able to use an age-0 recruitment index from the coastal sea grass surveys. This survey gave very low indices for 2011 and 2012, which provide almost all information (other than discards) on very recent recruitments. Low age-zero indices from this survey in the late 1990s were observed to produce substantial negative residuals from the SS3 fit, so in the absence of any other corroborating evidence, the 2011 and 2012 recruitment estimates from SS3 should be treated with some caution. This has implications for projections.

A major issue identified by the Review Panel was the evidence in support of the recent apparent very steep increase in SSB of Gag following the introduction of measures to reduce exploitation. The SS3 result indicates this is caused by the fish from some recent strong year classes attaining maturity, and experiencing a sharp reduction in fishing mortality. An increase is apparent in the headboat index (which is predominantly young fish) but less apparent in the other recreational indices. The commercial indices are predominantly older Gag and would not be expected to show this increase yet. Of the fishery-independent indices, the SEAMAP Video has a selectivity that should allow the steep increase in biomass in recent years to be observed. However, the observations at least in 2011 and 2012 are well below the expected values. The PC Video index was not well fitted, but has a large CV and short time span.

At the other end of the assessment series – the historical biomass trends prior to 1963 – the SS3 estimates are highly uncertain and are very sensitive to a range of input parameters such as stock-recruit steepness and initial log recruit offset, and to how the input data are weighted. A wide range of sensitivity runs carried out by the assessment team showed that the trends in biomass and recruitment over the data-rich part of the assessment are far less sensitive to model settings than is the case for the pre-1963 biomass and recruitment trends.

2. Exploitation estimates

I agree with the Panel conclusion that the different selectivity patterns among fleets are evident from the composition data and are appropriately estimated within the current model, and that given the large number of selection parameters estimated for the many fleets and abundance indices, further work should be focused on fixing selectivity parameters that are badly estimated from the data, or are highly correlated with other model parameters. The estimation of selectivity parameters may not be helped by treating catches as exact when they are measured with known error, as is the case for recreational catches. Some of the recreational catch estimates vary widely from one year to the next, which may be more a reflection of estimation error. Treating these as exact forces additional error into other estimations within the model, and to an extent defeats the purpose of using a statistical model.

b) Is the stock overfished? What information helps you reach this conclusion?

The SSB is estimated to have increased dramatically due to strong 2006-07 year classes entering the spawning stock (initially as females), combined with large reductions in exploitation rate on these year classes due to management measures. The perception of whether the stock is overfished depends, however, on how the SSB is calculated. The SSB-female model indicates that the stock is no longer overfished in relation to any of the proposed reference points. However, the SSB-combined (male plus female) model indicates that the stock is overfished in relation to SSBSPR30%, but the SSB is marginally above SSBMSY. These different perceptions reflect the greater time delay before the 2007-07 year classes enter the mature male population, as well as the different reference points for the two SSB metrics.

The recent rapid increase in biomass is indicated particularly by the headboat index (AW report Fig 3.2.13). Sensitivity runs were performed by down-weighting or leaving out the headboat index, and it was concluded that the recent steep biomass increase is robust to those changes. This reflects that the trends are heavily driven by the catch data composition indicating strong 2006-07 year classes and the weight of evidence for a steep decline in F following the low IFQs.

c) Is the stock undergoing overfishing? What information helps you reach this conclusion?

The stock does not appear to be undergoing overfishing. Recent management measures aimed at reducing the exploitation rate appear to have successfully lowered recent F values to below F_{msy} and also F_{spr30} for the agreed base model. The IFQs in particular have caused a large increase in discard rate across the size range, and the benefits of the measure in part depend on survival rates of discards. However the estimated reduction in F including assumed discard mortality is very large and unprecedented for the time series.

d) Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?

The shape of the stock recruit relationship is poorly defined, with very large recruitment variations at relatively low SSB and little information on the shape of the curve at high SSB. An asymptotic Beverton-Holt curve is assumed in fitting the model, and steepness is estimated by the base model to be high at >0.9 due to the lack of any information on the relationship between recruitment and SSB towards the origin or at high SSB. The Review Panel recommended that a fixed value of 0.85 (closer to values derived from meta-analyses) be used for the base case, with results also to be provided for management using the estimated high value and a lower fixed value of 0.7.

The interpretation of the stock recruitment relationship is also affected to some extent by the decision of the assessment team to input a low value of initial recruitment offset to better fit the early fishery data, which leads to an unusual run of very low recruitment deviations in the 1960s. The Review Panel noted from sensitivity analysis that this assumption had relatively little impact on stock trends in the data-rich period of the assessment, but without any *a priori* reason for such low recruitment (e.g., environmental conditions) it suggests that the early fishery data may have low information content or biases affecting estimation of pre-1970 recruitment. These recruitment estimates should not be plotted on a stock-recruit plot if they are considered to have little basis in reality. I agree with the Review Panel that this problem may also interfere with model estimates of steepness, and it would be preferable if a configuration could be found where the issue could be avoided. Setting the initial recruitment offset so that early recruitments are more in line with the stock-recruit model predictions is not too different to fixing steepness at a biologically more appropriate value in that both lead to a worse fit to the model, but may be a more reasonable decision if there are any concerns about the quality (bias / precision) of the early catch and composition data.

e) Are the quantitative estimates of the status determination criteria for this stock reliable? If not, are there other indicators that may be used to inform managers about stock trends and conditions?

In general, the estimated relative trends in biomass, recruitment and fishing mortality over the data-rich period of the assessment were robust to a wide range of sensitivity analyses, and the jitter analysis showed there was a well defined solution. However, uncertainties in parameters that are related to productivity, such as natural mortality and stock-recruit steepness, will affect stock status relative to reference points. The SSB metric that is adopted (females or combined-sex) also has a major impact. The Review Panel recommended that six alternative outcomes on status should be presented to managers: SSB (female only), or SSB (females+males), each with three steepness options (as estimated; 0.85 and 0.7).

4. Evaluate the stock projections, addressing the following:

a) Are the methods consistent with accepted practices and available data?

The methods for stock projections were consistent with accepted practices and available data.

Deterministic projections were run to evaluate stock status and associated yields for a range of fishing mortality rate scenarios: Fcurrent (fishing mortality rates for all fleets set to the geometric mean of the past three years, 2010-2012), Fspr30% (the fishing mortality rate that results in an equilibrium SPR of 30%), Fmax (the fishing mortality rate that maximizes the yield-per-recruit), and Foy (75% of FSPR30%). The Panel recommended that the base case should be the option with combined-sex SSB and steepness fixed at 0.85, with all other parameters set as the Base run from the Assessment Report. Benchmarks for the SPR 30% reference point and projections for the base model are presented in Table 3.2.8 of the Assessment Report. Benchmarks for the SPR 30% reference point and projections for the fixed steepness model are presented in Table 3.2.9 of that report.

b) Are the methods appropriate for the assessment model and outputs?

The bootstrap projections and the deterministic equivalents were internally consistent in carrying forward the population estimates, selectivity and retention parameters and catch allocations between fleets from the historical assessment phase. The assumption that selectivity, retention and catch allocations were the same as estimated for the three most recent years (2010-2012) was appropriate given the changes in management measures in that period. Forecast recruitments (deterministic or bootstrap) were derived from the model estimated Beverton-Holt stock-recruitment relationship, based on the recent time period (i.e., 1984-2011).

Uncertainty in stock status and forecasted yields for the projection years was investigated using the bootstrap approach. Random recruitment deviations (log values) for the projection period were created from a normal distribution with mean of 0 and standard deviation equal to the model estimated standard deviation of log recruit residuals for 1984-2011. This is a widely adopted approach for projections. The bootstrap datasets were not available during the review workshop but were sent to the Panel shortly thereafter.

c) Are the results informative and robust, and useful to support inferences of probable future conditions?

The projections are informative, and robust (bracketing a range of model configurations), provided that the underlying assumptions regarding future selectivity and retention patterns remain the same as estimated for the last three years. The two estimated very weak year classes in 2011 and 2012, estimated primarily from the age-0 sea grass habitat survey, were carried forward into the projection, although the robustness of these values is poorly known (previous similar indices were interpreted as large negative residuals in SS3). The impact of these year classes is likely to be greatest on recreational catches and discards in the short term.

d) Are key uncertainties acknowledged, discussed, and reflected in the projection results?

The projections bracket options for the SSB metric and stock-recruit steepness, and the bootstrap projections propagate the uncertainties in estimates and process error in recruitment in an internally consistent way.

5. Consider how uncertainties in the assessment, and their potential consequences, are addressed.

(a) Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods; (b) Ensure that the implications of uncertainty in technical conclusions are clearly stated.

The Review Panel report provides a clear analysis of how uncertainties in the assessment are represented and estimated, and reasons why the model estimates of uncertainty can be unreliable. The total uncertainty is the combination of estimation error within the model and additional uncertainty regarding model structure, input parameters and assumptions, normally evaluated through sensitivity analysis and scenarios. Implicit in fitting the SS3 model is the assumption that errors in the data are random and independent (unless serial correlation is built in). This is almost never the case, and data sets have some form of bias related to design, implementation and analysis which may vary with time. In addition, key parameters such as natural mortality and true shape of the stock-recruit curve are seldom known and may indeed drift over time due to changes in environment and predator abundance. Sources of uncertainty have been treated in a variable way within the Gag grouper assessment:

- Relative precision of different data sets is input to SS3 as CVs or (for composition data) effective sample sizes or proxies. In some cases, for example the MRFSS/MRIP surveys, the data are collected using statistically sound designs aimed at minimizing bias and obtaining robust estimates of precision. Unfortunately, for recreational fisheries the precision estimates of catch quantities have been ignored in the base model and the catches are treated as exact. The true precision of the composition data for recreational and commercial fisheries, and how this has changed over time, is not reflected due to the use of numbers of fish measured or aged as proxies for effective sample size, and capping these at 200 to avoid over-weighting where many fish were measured. For relative abundance indices, precision is estimated based on fits to a delta-lognormal model. The net result is that the fishery data are heavily weighted relative to abundance index data, a possible reason for the roughly similar residual patterns across many of the abundance indices in recent years (see Fig. 1 above).

- Possible biases in historical catch estimates are likely. This is due, for example, to the need to hind cast historical species compositions for mixed species catches based on more recent data. Recreational estimates in the early 1980s may have biases due to the evolution of the survey design and coverage. Hindcasting of pre-1980s recreational catches uses methods that are difficult to accurately ground truth. Sensitivities to alternative plausible scenarios for biased time series were not explored, although these may affect mainly the early stock trends.
- Fishery-dependent abundance indices (used extensively for Gag) have, by their nature, no design base. Modeling to standardize the series can only work up to a point, leaving short-term changes or longer term trends in catchability due to unknown factors such as fisher behavior and unaccounted-for changes in technology or efficiency. For example in recreational fisheries, it is hard to imagine that the widespread use of GPS and affordable high-resolution echo sounders has not had a significant impact on the ability to target seabed features associated with Gag hotspots. The guild approach to filtering trips may also have significant bias which has not been explored – for example where poor initial catch rates of Gag on a trip result in boats shifting to target different species (or vice versa). This may not be an issue but it is not proven in the material supplied to the review.
- An additional model uncertainty exists, caused by decisions on model configuration, such as shapes of selection curves, fixing of parameters that cannot be reliably estimated in the model (such as natural mortality or form of stock-recruit model), inclusion/exclusion of data, and how to weight individual data series.

Many of these issues may contribute to unusual residual patterns that are presented in detail in the assessment results but can be hard to explain because biases in the data, and the true relative quality of data sets, are poorly understood. The residual plots and plots of fitted lines and observations (Fig. 3.2.1-3.2.48) indicate that some fits are close to the observations (relative to the input precision of the data) and others are far off, in some cases systematically so. Some of the plots of the fitted indices also show auto-correlated errors (periods of only negative residuals followed by periods of positive residuals), which is in contrast with the assumed independent error structure.

The assessment team did a good job in exploring the sensitivity of the model to a wide range of uncertainties related to model configuration and input parameters, both before the review meeting and during it, which greatly helped the review process. The overall impression was that the model results are more robust over the recent period than in the first ~12 years of the "data rich" period. However, the parametric bootstrap or inverse Hessian estimates of confidence intervals will, overall, underestimate the true confidence intervals, perhaps substantially. The evaluation of uncertainty in assessment outputs is very complex issue for this type of assessment, and it is difficult to express the true uncertainties in estimates given all these issues in combination. However, to facilitate this, it is important that the model is given accurate information on the relative precision of the data sets, and that biases in the data are well understood from direct investigation of data quality external to the model.

I agree with the conclusions of the Review Panel that the uncertainties in the assessment are not sufficient to invalidate the results as a basis for providing informed management advice, and that the robustness of the assessment has been adequately explored and the implications of uncertainty clearly stated in all relevant graphs and tables. A critical uncertainty in the perceived stock status of Gag grouper identified by the assessment team is the decision on whether to use female-only SSB or combined-sex SSB for defining biomass reference points and status relative to these, which is not related to many of the uncertainties inherent in the assessment model.

6. Consider the research recommendations provided by the Assessment workshop and make any additional recommendations or prioritizations warranted.

a) Clearly denote research and monitoring that could improve the reliability of, and information provided by, future assessments.

The Review Panel collated the many research recommendations made by the Data and Assessment workshops, and made some additional recommendations. The recommendations by the Data and Assessment workshops are all sensible ideas that would help improve the ability to assess the stock in the future. However no attempt has been made to evaluate the relative impact of the additional or new knowledge gained, or the feasibility or cost-effectiveness of the proposals. As a reviewer and non-expert on these species, it is therefore difficult for me to make sensible comments on these proposals. I will therefore make only some general comments.

- The protogynous hermaphrodite life history trait appears to be a major issue in determining the biomass or spawning potential metric to use for defining reference points and stock status relative to these. Well-targeted research is needed to improve knowledge of how the transition is triggered, and the impact of over-depletion of the mature male population on productivity and genetic diversity.
- The assessment is currently dependent on fishery-dependent abundance indices, and it is known that fisher behavior has been affected strongly by the reduced IFQs. The indices show varying degrees of catchability trends, strongest in the commercial fleets. The only fishery-dependent survey covering an extensive part of the along-shore stock range is the SEAMAP video survey, although this does not support the recent large biomass increase shown by the assessment. Other video surveys target smaller parts of the range of the stock and different age groups. In the longer term, the ability to give reliable advice on this stock would be greatly enhanced by establishment of an integrated fishery-independent survey that is based on sound statistical design, covers the range of the stock and as many components of the stock as possible, provides indices or estimates of abundance where the relationship with true abundance is well understood, is cost-effective and has good precision, and provides information on as many species as possible along with relevant habitat and other environmental information. It is possible that the existing video surveys could be further developed and enhanced to provide the necessary stratified random observations, and include collection of fish for biological analysis. A standardized

approach for the video survey and fish sampling, as adopted for many trawl surveys, would be needed. Other approaches for large-scale fishery-independent surveys could also be considered.

- Consideration should be given on how best to allocate resources for collecting length and age samples from the commercial and recreational fisheries. To the greatest possible extent, such collections should follow sound statistical, probability-based design and avoid over-sampling of clusters and focusing on numbers of primary sampling units needed to achieve a desired precision.

(b) Provide recommendations on possible ways to improve the SEDAR process.

The Review meeting was productive, and the assessment team appreciated the advice given by Panel members with extensive experience in the application of statistical assessment models including Stock Synthesis. I agree with the following Review Panel recommendations to improve the SEDAR process:

> (1) Due to the inherent complexity of highly parameterized statistical catch at age models (i.e. Stock Synthesis) and the relative scarcity of expert users, the review panel recommends that each SEDAR assessment workshop panel include at least one nationally recognized expert in the model used (e.g. Stock Synthesis). This expert could participate in person or by electronic means and would greatly facilitate the review process.
>
> (2) There is concern over a variety of issues that emerge as a result of the Assessment Workshop being exclusively performed via webinars. The Review Panel emphasizes the importance of face-to-face meetings for improving the model development during the assessment phase. The panel feels that many of the issues uncovered during the review process could have been avoided, and this may have enabled the assessment team to provide a more polished product for review, resulting in the best model possible.

In addition, I have the following comments on the SEDAR process:

- A better framework for documenting the quality of data sets, in an easily accessible way, would facilitate the assessment and review processes. This is necessary to ensure that the precision metrics input to Stock Synthesis (CVs of catches and abundance indices; effective sample sizes for composition data; age error CVs etc.) correctly represent the relative precision of input data. It is also needed for identifying biases in a way that could help decisions on inclusion or additional weighting of data sets, and to help interpret residual patterns in the model. Such a framework would also help to identify where work is most needed through improved design, implementation and analysis to improve the quality of data where it is most needed to improve the assessment. A number of ICES expert groups have worked towards implementation of a quality assurance framework for fishery and biological data, in some cases involving US experts. Their reports are available on the ICES website[2] by searching on acronyms PGCCDBS, WKPICS, SGPIDS, WGRFS).

7. Provide guidance on key improvements in data or modeling approaches which should be considered when scheduling the next assessment.

The Review Panel considered that for Gag, the Stock Synthesis modeling framework remains appropriate for the type of data available, allowing flexibility to account for changes in size limits or IFQs that affect patterns of discarding in commercial and recreational fisheries. Currently the model structure and implementation appear appropriate, although further work is needed to resolve the issue of poor definition of the initial slope of the stock recruit curve. The Review Panel suggested the following work for improving the assessment:

(1) Research should be conducted for the most appropriate value of steepness to be used for Gag Grouper – either across a range of species (e.g. Ram Myers database) or through use of a well-estimated value from a closely related stock or species.

(2) If an appropriate fixed value for steepness is found, further research to explore the estimation of parameters currently fixed in the model, such as natural mortality.

(3) Further work on improving selectivity parameters that are poorly estimated from the data available, or highly correlated with other model parameters.

Future improvements are also likely to be achieved through improvements in key data sets and in the understanding of the biology. A key aspect of biology where there is currently only limited understanding is what determines the probability of transition from female to male, how far the male population can be depleted before sperm limitation starts to impact productivity, and if reduced numbers of males would lead to transitioning at a lower size with impact on population egg production. Further work is needed to resolve these uncertainties, which impact mainly the choice of biomass reference points and monitoring of biomass relative to these.

Ensuring the continued quality of length and age compositions for retained and discarded fish is important for fitting year class strength, selectivity and retention.

Currently, the most influential relative abundance indices are from recreational and commercial fisheries, i.e. the same data sets used for estimating catch compositions and recreational catches, but filtered using information on species guilds in catches to try and identify trips where gag grouper have a probability of being caught. Further work is needed to identify potential biases in these approaches, for example where Gag were initially targeted in a recreational trip but zero or low catch rates led to a switch to other areas or methods that do not catch this species. Other factors affecting catch rates in hook fisheries, particularly longlines (e.g. gear saturation, competition with other species) should be considered in evaluating if the commercial index series are reliable. Further investigation into the robustness of the design of the video surveys should also be carried out in relation to coverage of the stock and density-dependent selection of habitats.

[2] http://www.ices.dk/publications/our-publications/Pages/Expert-Group-Reports.aspx

4.2 Greater Amberjack

My evaluation of the assessment and review process are given below by Term of Reference for the review process. This includes many outcomes documented in the Review Panel report, which was compiled collaboratively by the Panel, expanded where necessary with my own more detailed opinions and recommendations. Many of the issues are common to the Gag grouper and amberjack assessments, and a large amount of text from my Gag grouper review are replicated below, so that the amberjack text can be read in isolation if needed.

ToR. 1. Evaluate the data used in the assessment, addressing the following:

a) Are data decisions made by the Assessment Workshop sound and robust?

With some exceptions, the decisions made by the Data and Assessment Workshops were sound and robust, and the data used in the assessment are adequate and appropriate for that purpose. The Stock Synthesis modeling framework was chosen to allow incorporation of a range of fishery and abundance index data by length and age, in a way that can account for changes in fishery data caused by management actions affecting selectivity and retention (discarding). This is not possible for the continuity ASPIC assessment. The Data Workshop conducted an extensive compilation and evaluation of all available fishery dependent and independent data sets, and reviewed methods for determining input values for natural mortality and discard mortality. Decisions were made on how to make imputations where data were missing or inadequate; approaches for extrapolating data series back in time; and inferring historical amberjack catches from mixed-species catch records. Fishery-dependent abundance index data were filtered using species guild criteria in an attempt to remove trips in areas beyond amberjack habitat. All fishery-dependent and independent abundance data were analyzed using delta – lognormal modeling to remove the effect of factors other than abundance affecting interannual changes in observations such as fishery catch rates or video camera counts. Overall, the raw data were subject to extensive processing to provide assessment inputs, possibly to a greater extent than for Gag grouper.

All decisions were well explained in the workshop reports and presentations to the Review Panel. In most cases the decisions were sound and robust, and where appropriate their effect on the assessment was adequately explored in sensitivity analyses. However, in my opinion, the Assessment Workshop made two inappropriate decisions (as also done for Gag grouper):

- Firstly, recreational catches were treated as exact, when in fact they are estimated from probability-based surveys which are designed to minimize biases as far as possible and to provide robust estimates of precision. This is in contrast to commercial landings figures which are based on census (i.e. very high precision) but may have biases due, for example, to problems with species reporting. Statistical models should be fitted taking into account the known precision of the input data (where random measurement error is the main source of uncertainty), whilst biases should be explored through sensitivity analysis.

- Secondly, a decision had been made to use numbers of fish measured or aged, capped at 200, as input effective sample sizes (ESS) for fishery composition data. It is well known that numbers measured or aged is a poor indicator of relative precision due to cluster sampling effects. The use of individual fish numbers resulted in the composition data for many fleet-year combinations being capped to avoid very high

> weighting of some years and fleets. Stock Synthesis generates output ESS values based on the fit to each year's data, but extensive capping of input ESS precludes the ability to examine the correlation between input and output ESS. If poorly correlated, but the input ESS correctly reflect the relative precision of the input data, it would suggest that factors other than random sampling error are degrading the model fit to the data. The Review Panel strongly recommended that if the true ESS is not calculated, a proxy should be developed, for example the number of independent primary sampling units (such as numbers of trips sampled). During the Review Meeting, it was commented that sampling in the earlier years tended to involve fewer trips and more fish measured per trip than in later years.

Additional assessment model runs were requested at the Review Meeting to treat the recreational catch estimates as subject to survey error rather than being exact, and using numbers of trips sampled for length or age as proxies for effective sample sizes, without any capping. The assessment team considered that for this species it was not possible (at the time of the RW) to determine the historical number of fishing trips where amberjack were sampled, and they were unsure which values for precision of recreational catch estimates should be used. Further work is needed to identify the most suitable model inputs on effective sample size for composition data and recreational survey precision.

b) Are data uncertainties acknowledged, reported, and within normal or expected levels?

Data uncertainties were explored and reported, although it would have been more helpful to see a clearer summary of the relative quality of the different data sets. Where they could be reliably quantified, uncertainties appeared to be within normal or expected levels given the design of data collection schemes and the amount of sampling that has taken place. Fishery data limitations were due, in part, to greater amberjack not being a major directed fishery and therefore present in a relatively small fraction of trips. As mentioned under ToR1(a), the raw data were subject to extensive processing to provide assessment inputs, and it is not clear how much of the final estimates of catches or compositions is the result of imputation. Many of the approaches adopted, such as imputations for missing data for years, fleets or areas, corrections for mixed-species reporting, filtering of relative abundance data etc. can themselves lead to variable bias over years. The Data and Assessment Workshops acknowledged the major uncertainties and described how they dealt with these. However, it was difficult as a reviewer to clearly understand the relative quality of the different data sets, either from the workshop reports or during the review process. In some cases, for example the recreational fisheries (which take a large fraction of the catches), the data are collected using statistically-sound sampling designs to reduce bias and allow more accurate estimates of precision, and the quality of the data are easier to evaluate. In such cases, uncertainties appeared to be within normal or expected levels given the design of data collection schemes and the amount of sampling that has taken place. In other cases, for example commercial fishery sampling at sea or on shore, filtered relative abundance data from commercial fleets, and imputation or hindcasting of missing data, it is not clear what types and magnitude of bias might be present.

In general, a clearer framework for documenting known or potential data quality issues (bias and precision) in relation to design, implementation, sampling achievement and analysis of data over different periods, using suitable quality indicators, would be very helpful for assessment analysts and reviewers. Evaluating data quality through performance in an assessment model is not sufficient in itself if the errors in the data include biases and variance. Although the relative abundance series were subject to detailed statistical modeling, I felt that more consideration could have been given to *a priori* evaluation of whether the data are capable of providing indices directly proportional to abundance. For example, a number of factors such as area were included in the commercial longline and handline index computation (Saul: AW18), but no discussion was included on potential effects of other factors such as competition and saturation of hooks by competing species (which may have different trends in abundance to amberjack), or technological changes that may have altered catchability over time. The AW18 report noted that the indices for the commercial handline and longline indices diverged and that this disagreement could be an artifact of changes in fisher behavior, such as the longline fleet being forced to fish further offshore in more recent years in response to changing gear based regulations.

c) Are data applied properly within the assessment model?

Based on the workshop reports and presentations at the Review Workshop, the data were properly applied within the ASPIC and Stock Synthesis assessment models.

Deficiencies and uncertainty in the data were explored. Consideration was given to appropriate fitting of data to SS3 and, for comparison, to ASPIC Version 5.34 which was used to fit a non-equilibrium production model for continuity with the 2010 stock assessment update. Some of the fishery-dependent indices (MRFSS and the commercial handline) developed during SEDAR 33 and recommended in the SEDAR 33 DW differ in trend from the indices from previous update assessment. Because ASPIC model results can be sensitive to changes in the indices, the methods used to develop the SEDAR 33 indices for Greater Amberjack were explored in depth during various SEDAR 33 assessment workshop webinars. The indices were recomputed in terms of weight to be consistent with the basis of production models where the population dynamics are in terms of biomass rather than numbers. Previous SEDAR 9 and SEDAR 9 Update assessments for Greater Amberjack used indices developed in numbers per unit effort which only makes sense if the average weight of individuals remains constant on average over time, even with the imposition of size limits. The SEDAR 33 assessment panelists reviewed and rejected that assumption.

d) Are input data series reliable and sufficient to support the assessment approach and findings?

I agree with the conclusion of the Review Panel that the input data series are not, at present, of sufficient quality to support a stable and reliable implementation of the proposed Stock Synthesis model at the level of complexity and parameterization set up by the SEDAR 33 Assessment Workshop. The Panel agreed that Stock Synthesis was an appropriate modeling framework to deal with changes in size structure of catches or fishery-dependent indices of abundance following changes in fishery selectivity or retention, and to allow inclusion of different types of length and age data for variable blocks of years. However, during the course of the review, when viewing the diagnostics and the sensitivity of the SS3 model solution to jitter in starting values, the Panel concluded that the base model configuration was over-parameterized given the quality of the composition data in particular, and that a number of other decisions related for example to data weighting could be improved. The nature of greater amberjack and its fisheries means that data quality is patchy and in places insufficient to support the estimation of the many selectivity and other parameters in the SS model. The Panel considered that further development could be carried out after the review meeting to identify a simpler Stock Synthesis model formulation appropriate to the information content of the available data. At the same time, well-targeted improvements in the data are needed in the longer term to ensure the model can reliably account for changes in selectivity or discard practices following changes in size limits and IFQs. The current data series are also problematic for input to the continuity ASPIC model, which cannot account for changes in size structure of catches or fishery-dependent indices of abundance following changes in fishery selectivity or retention.

2. Evaluate the methods used to assess the stock, taking into account the available data.

a)-c): Are methods scientifically sound and robust? Are assessment models configured properly and used consistent with standard practices? Are the methods appropriate for the available data?

Continuity assessment: ASPIC

The model used in the previous benchmark - ASPIC (A Stock-Production model Incorporating Covariates) - was configured as the continuity assessment model for amberjack. Whilst ASPIC is statistically sound and robust, it is not an appropriate method where there are dynamic trends in fishery selectivity or retention. It does not use fishery or survey composition data that can provide information on changes in selectivity or retention, or in recruitment which is a major component of annual population growth. (At the same time it is noted that the composition data for amberjack are of poor quality for some fishery and year combinations.) Changes had been made by the assessment team to convert recreational fishery catch numbers to weights, which is the appropriate configuration for production modeling. However, this involved uncertainty in choice of mean weights to apply, necessitating two scenarios: the “High Case” which assumed that recreationally discarded fish had the same average annual size as all recreationally landed individuals, and the “Low Case” which assumed that discards had the same average size as fish that were below the size limit and landed prior to 1990 when size and bag limits were introduced. This assumption affected the estimates of annual catch weight for each recreational fishery and the indices of biomass for the combined recreational fishery. The conversion to weights generated different results than given by the previous benchmark assessment. On balance, the Review Panel unanimously agreed that ASPIC was no longer a suitable approach for assessing the status of Gulf of Mexico Greater Amberjack.

Stock Synthesis

The Stock Synthesis 3 (Methot 2013) modeling framework provided as the SEDAR 33 proposed base method is widely used in the USA and elsewhere and is well tested in peer-reviewed assessments. Its strengths include the ability to fit to mixtures of length and age composition data that can include missing years, and to explicitly model the selectivity and retention patterns of component fisheries and any changes in these over time. This is important and appropriate given the nature of the amberjack data and the major changes in management measures affecting retention (discarding) practices. The model also allows the ability to explore important parameters such as those defining the shape of the stock-recruit relationship. The approaches adopted to model selectivity (constant over time) whilst allowing changes in retention ogives to account for changes in discarding practices, the effects of which were evident in the catch composition data, were in principle statistically sound and robust. However, poor fits to many of the fishery composition data series, and unstable solutions, indicated model identification problems that could not be resolved during the Review Workshop. The Review Panel members were all of the opinion that Stock Synthesis remained an appropriate assessment approach for this stock, but the optimal configuration of the model had not yet been found. The Reviewers' identified issues that should be addressed before the assessment model could be accepted as properly configured and consistent with standard practices. The main problem identified by the Review Panel is given below, based on the Panel Report:

The Review Panel's main concern was the so-called jitter analysis. In a jitter analysis it is the intention to verify model convergence by starting all the model parameters at numerous different initial values (within some range) and then see if the end result in terms of the objective function, estimated model parameters, and important output metrics is unchanged. A detailed look at the results showed that for the individual likelihood contributions of catch, survey and the length and age compositions (AW report figure 3.2.2.1), the model did not converge to one unique solution, and important metrics such as ratio of current F to reference F levels were also changing (e.g. F_{2012} /F $_{SPRTtgt}$ varied by about 10% when the starting point was changed by 10%). Another place where the convergence problem was evident was the profile likelihood with respect to the steepness parameter (AW report Fig. 3.2.4.1), where sudden inexplicable high values occur in several places on otherwise convex curves. The Review Panel noted that occasional lack of convergence could be solved in each individual case by choosing different starting points until convergence, but in the presented model the point of convergence is highly dependent on the arbitrary starting values and the model, in its current configuration, was not finding a unique minimum, which normally occurs if the model is non-identifiable (i.e. over-parameterized, where a change in some model parameters can compensate for a change in some other model parameters). To solve such an issue it is often necessary to fix some parameters, or assign priors to them. When looking for which model parameters to restrict it can be useful to look at correlations between model parameters, and to see if the standard deviations from a parametric bootstrap are similar to those derived from the inverse Hessian approximation. If they are very different it could be an indication of over-parameterization. Some of these methods were tried during the review meeting, and some results were improved. Some selectivity parameters were identified as problematic ones, which means that the Review Panel and the assessment team were optimistic that a more parsimonious and stable SS3 model configuration could be developed.

In conclusion, the consensus view of the Review Panel is that relatively minor adjustments to the Stock Synthesis configuration would make it identifiable and suitable as basis for management of Greater Amberjack. In the course of doing this, I recommend that other improvements to the model should be made, including inputting more appropriate measures of effective sample size for composition data (either true ESS estimates are a proxy related to number of primary sampling units sampled, typically numbers of fishing trips sampled), and inputting precision estimates for annual recreational catch and discard numbers which are derived from probability-based surveys.

3. Evaluate the assessment findings with respect to the following:

A) Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?

Amberjack ASPIC

Given the dynamic changes in fishery retention patterns which affect the trends in landings and discards weights as well as the fishery-dependent abundance indices, the consensus view of the Review Panel is that ASPIC provides continuity with previous assessments, but is no longer the preferred method for estimating trends in biomass and exploitation, and determination of current stock status relative to production model estimates of reference points such as F_{MSY} and B_{MSY}. To deal with changes in retention that affect the continuity of the data series, the series need to be broken into segments which adds additional catchability parameters to be estimated and reduces the data contrast needed to estimate the stock productivity parameters.

Amberjack Stock Synthesis

The consensus view of the Review Panel was that the presented configuration of the SS3 model was not acceptable as a basis for advice, and that a more parsimonious configuration appropriate to the data quality was needed, along with some other amendments to allow more correct weighting of data sets according to their quality.

Despite issues with the model configuration and fit to composition data, the fit to the selectivity-adjusted fishery abundance indices was reasonable, suggesting that the reconfigured Stock Synthesis should be capable of providing information on trends in biomass and exploitation rate. Of the fishery-independent indices, PC Video and Seamap Video both had observations that were variable but mostly flat, with expected fits also showing no strong trends.

As I have commented earlier, I felt that more consideration could have been given to an *a-priori* evaluation of whether the fishery-dependent and independent data are capable of providing indices directly proportional to abundance, referring to the example of commercial longline and handline index computation (Saul: AW18) where there is no discussion on potential effects of other factors such as competition and saturation of hooks by competing species (which may have different trends in abundance to amberjack), or technological changes that may have altered catchability over time. It is interesting that where such gears are used in design-based surveys (e.g., Alaska sablefish longline survey), considerable efforts are made to investigate factors such as hook competition and saturation, but this is not considered for non-design based commercial CPUE indices. The Assessment report shows the longline index increasing at a faster rate than the commercial handline index over the series.

B) Is the stock overfished? What information helps you reach this conclusion?

Because a base model configuration was not identified during the Review workshop, the Panel was unable to make this determination. Continuity runs using ASPIC found the stock to be at or slightly above the overfished status, however the model is less appropriate where the catch and abundance index data are affected by large changes in retention, as is the case for amberjack. The current ASPIC assessment shows substantial improvement of stock condition compared with the 2009 update - although this improvement was likely caused by changes to the calculation of the indices. The Stock Synthesis model runs were highly variable in their evaluation of the stock condition although most runs indicated an overfished stock

C) Is the stock undergoing overfishing? What information helps you reach this conclusion?

Because a base model configuration was not identified during the Review workshop, the Panel was unable to make this determination. While showing highly variable results with regard to overfished status, ASPIC and most SS runs presented to the panel did not indicate current overfishing

D) Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?

Because a base model configuration was not identified during the Review workshop, the Panel was unable to make this determination. The SS3 runs presented suggested that the shape of the stock-recruitment relationship is poorly characterized. In such a situation it is best to investigate the effect of a plausible range of steepness values, perhaps informed by similar species or a meta-analysis

E) Are the quantitative estimates of the status determination criteria for this stock reliable? If not, are there other indicators that may be used to inform managers about stock trends and conditions?

As a base model configuration was not identified during the Review workshop, the Panel was unable to make this determination.

4. Evaluate the stock projections, addressing the following:

a) Are the methods consistent with accepted practices and available data?

The methods for stock projections were consistent with accepted practices and available data. However, because a base model configuration was not identified during the Review workshop, the Panel was unable to evaluate the projections.

b) Are the methods appropriate for the assessment model and outputs?

The projection methods were appropriate for the SS3 model and outputs, but because a base model configuration was not identified during the Review workshop, the Panel was unable to evaluate the projections

c) Are the results informative and robust, and useful to support inferences of probable future conditions?

Because a base model configuration was not identified during the Review workshop, the Panel was unable to evaluate the projections.

d) Are key uncertainties acknowledged, discussed, and reflected in the projection results?

Because a base model configuration was not identified during the Review workshop, the Panel was unable to evaluate the projections.

5. Consider how uncertainties in the assessment, and their potential consequences, are addressed.

a) Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods; b) Ensure that the implications of uncertainty in technical conclusions are clearly stated.

ASPIC continuity model

For the ASPIC model, the important uncertainties are less to do with model fit and more to do with the effect of changing retention patterns on the catches and abundance indices and how that was handled by segmenting the index data. The two options to convert discard numbers to weights added a small amount of additional uncertainty explored through sensitivity runs. Figure 2 below shows some undesirable features in the relative abundance series as inputs to ASPIC – firstly the need to segment some series at the beginning and end of the series; secondly the transient high indices for combined charter and private recreational fisheries in 1991 and 1992, immediately after the increase in size limit (CBPR Low and High – the two options for estimating discard weights), not seen in such a pronounced extent in other indices; and thirdly the very different trends indicated by the commercial long line fleet compared with the other fleets, which may partly reflect differences in selectivity and retention.

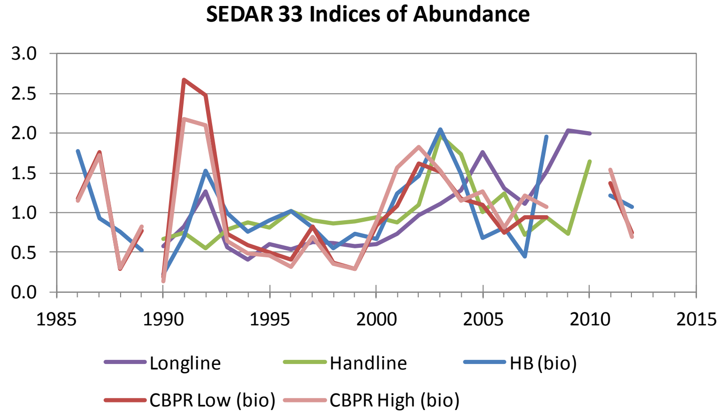

Fig. 2: Mean-standardized abundance indices (by weight) used in the ASPIC base run (from Review Panel presentation)

For the ASPIC model a bootstrap method with 1000 runs was used to compute confidence intervals and is a reasonable way to ensure that the uncertainties are correctly propagated in a non-linear model. The AW report clearly states the uncertainties with respect to important output metrics (AW report Fig. 3.4.3.1 and Table 3.3.5.1.). In addition, sensitivities to B1/K input ratio, discard mortality, and index weighting were conducted for the ASPIC results. These showed expected differences, but also that most conclusions about important output metrics were relatively stable (AW report Fig. 3.4.6.1.2, 3.4.6.2.1). The important output metrics from the ASPIC model were however sensitive to the new data compilation in SEDAR33 compared to the SEDAR9 update (AW report Section 2.6.1). When 2009 is considered to be the final year and the new indices are used, the estimate of B/BMSY is doubled, and F/FMSY estimate is halved (AW report Fig. 3.4.7.1 and Table 3.4.7.1).

Stock Synthesis

The total uncertainty in the Amberjack SS3 assessment is the combination of estimation error within the model (quantified by parametric bootstrap or Hessian) and additional uncertainty regarding model structure, input parameters and assumptions, normally evaluated through sensitivity analysis and scenarios. The Review Panel concluded that the Hessian standard deviations for Amberjack are not likely to be reliable because the model with the present configuration is not identifiable, but the standard deviations from the 1500 parametric bootstrap include all aspects of the implemented model when propagating uncertainties from observations to estimates of model parameters. This means that even if this model is not strictly identifiable (as some model parameters can compensate for others) this will, to some extent, be captured in the bootstrapped uncertainties.

However, implicit in fitting the SS3 model is the assumption that errors in the data are random and independent (unless serial correlation parameters are included). This is almost never the case in such assessments, and data sets have some form of bias related to design, implementation and analysis, which may vary with time. In addition, key parameters such as natural mortality and true shape of the stock-recruit curve are seldom known and may indeed drift over time due to changes in the environment and predator abundance. Sources of uncertainty have been treated in a variable way within the Amberjack SS3 assessment:

- Relative precision of different data sets is input to SS3 as CVs or (for composition data) effective sample sizes or proxies. In some cases, for example the MRFSS/MRIP surveys, the data are collected using statistically-sound designs aimed at minimizing bias and obtaining robust estimates of precision. Unfortunately, for recreational fisheries the precision estimates of catch quantities have been ignored in the base model and the catches are treated as exact. The true precision of the composition data for recreational and commercial fisheries, and how this has changed over time, is not reflected due to the use of numbers of fish measured or aged as proxies for effective sample size, and capping these at 200 to avoid over-weighting where many fish were measured. For relative abundance indices, precision is estimated based on fits to a delta-lognormal model.
- Possible biases in historical catch estimates are likely. This is due, for example, to the need to hindcast historical species compositions for mixed species catches based on more recent data. Recreational estimates in the early 1980s may have biases due to the evolution of the survey design and coverage. Hindcasting of pre-1980s recreational catches used methods that are difficult to ground truth. Sensitivities to alternative plausible scenarios for biased time series were not explored, although these may affect mainly the early stock trends.
- Fishery-dependent abundance indices (used extensively for Amberjack) have, by their nature, no design base. Modeling to standardize the series can only work up to a point, leaving short-term changes or longer term trends in catchability due to unknown factors such as fisher behavior and unaccounted-for changes in

technology or efficiency. For example in recreational fisheries, it is hard to imagine that the widespread use of GPS and affordable high-resolution echo sounders has not had a significant impact on the ability to target features where Amberjack aggregate. The guild approach to filtering trips may also have significant bias which has not been explored – for example where poor initial catch rates of Amberjack on a trip result in boats shifting to target different species (or vice versa). This may not be an issue but it is not proven in the material supplied to the review.

- An additional model uncertainty exists, caused by decisions on model configuration, such as shapes of selection curves, fixing of parameters that cannot be reliably estimated in the model (such as natural mortality or form of stock-recruit model), inclusion/exclusion of data, and how to weight individual data series.

Many of these issues may contribute to unusual residual patterns that are presented in detail in the assessment results but can be hard to explain because biases in the data, and the true relative quality of data sets, are poorly understood. The residual plots and plots of fitted lines and observations indicate that some fits are close to the observations (relative to the input precision of the data) and others are far off, in some cases systematically so. Some of the plots of the fitted indices also show auto-correlated errors (periods of only negative residuals followed by periods of positive residuals), which is in contrast with the assumed independent error structure.

The assessment team did a good job in exploring the sensitivity of the model to a wide range of uncertainties related to model configuration and input parameters, both before the review meeting and during it. The overall impression was that the model results are more robust over the recent period than in the early part of the "data rich" period. The evaluation of uncertainty in assessment outputs is a very complex issue for this type of assessment, and it is difficult to express the true uncertainties in estimates given all these issues in combination. However, to facilitate this, it is important that the model is given accurate information on the relative precision of the data sets, and that biases in the data are well understood from direct investigation of data quality external to the model.

A final issue related to uncertainty was the steepness parameter. The SS3 configuration proposed by the assessment team estimated the steepness, and the profile likelihood (vaguely) suggested that estimation was possible (AW report Fig. 3.2.4.1). Model improvements made during the review meeting however changed this perception, and the conclusion of the reviewers was that estimating steepness is very uncertain, and that fixed values should be chosen. If fixed values for steepness are used, it may be possible to explore the estimation other highly influential parameters such as natural mortality.

6. Consider the research recommendations provided by the Assessment workshop and make any additional recommendations or prioritizations warranted.

Clearly denote research and monitoring that could improve the reliability of, and information provided by, future assessments.

The Review Panel collated the many research recommendations made by the Data and Assessment workshops, and made some additional recommendations. The recommendations by the Data and Assessment workshops are all sensible ideas that would help improve the ability to assess the stock in the future. However no attempt has been made to evaluate the relative impact of the additional or new knowledge gained, or the feasibility or cost-effectiveness of the proposals. As a reviewer and non-expert on these species, it is therefore difficult for me to make sensible comments on these proposals. I will therefore make only some general comments.

- The assessment is currently heavily reliant on fishery-dependent abundance indices, and it is known that fisher behavior can be strongly altered by management measures for this or other species taken in the same fishery. The only fishery-dependent survey covering an extensive part of the along-shore stock range is the SEAMAP video survey. Other video surveys target smaller parts of the range of the stock and different age groups. In the longer term, the ability to give reliable advice on this stock would be greatly enhanced by establishment of an integrated fishery-independent survey that is based on sound statistical design, covers the range of the stock and as many components of the stock as possible, provides indices or estimates of abundance where the relationship with true abundance is well understood, is cost-effective and has good precision, and provides information on as many species as possible along with relevant habitat and other environmental information. It is possible that the existing video surveys could be further developed and enhanced to provide the necessary stratified random observations, and include collection of fish for biological analysis. A standardized approach for the video survey and fish sampling, as adopted for many trawl surveys, would be needed. Other approaches for large-scale fishery-independent surveys could also be considered.
- Consideration should be given on how best to allocate resources for collecting length and age samples from the commercial and recreational fisheries. To the greatest possible extent, such collections should follow sound statistical, probability-based design and avoid over-sampling of clusters and focusing on numbers of primary sampling units needed to achieve a desired precision.

(c) Provide recommendations on possible ways to improve the SEDAR process.

The Review meeting was productive, and the assessment team appreciated the advice given by Panel members with extensive experience in the application of statistical assessment models including Stock Synthesis. I agree with the following Review Panel recommendations to improve the SEDAR process:

> (1) Due to the inherent complexity of highly parameterized statistical catch at age models (i.e. Stock Synthesis) and the relative scarcity of expert users, the Review Panel recommends that each SEDAR assessment workshop panel include at least one nationally recognized expert in the model used (e.g. Stock Synthesis). This expert could participate in person or by electronic means and would greatly facilitate the review process.
>
> (2) There is concern over a variety of issues that emerge as a result of the Assessment Workshop being exclusively performed via webinars. The Review Panel emphasizes the importance of face-to-face meetings for improving the model development during the assessment phase. The Panel feels that many of the issues uncovered during the

review process could have been avoided, and this may have enabled the assessment team to provide a more polished product for review, resulting in the best model possible.

In addition, I have the following comments on the SEDAR process:

- A better framework for documenting the quality of data sets, in an easily accessible way, would facilitate the assessment and review processes. This is necessary to ensure that the precision metrics input to Stock Synthesis (CVs of catches and abundance indices; effective sample sizes for composition data; age error CVs etc.) correctly represent the relative precision of input data. It is also needed for identifying biases in a way that could help decisions on inclusion or additional weighting of data sets, and to help interpret residual patterns in the model. Such a framework would also help to identify where work is most needed through improved design, implementation and analysis to improve the quality of data where it is most needed to improve the assessment. A number of ICES expert groups have worked towards implementation of a quality assurance framework for fishery and biological data, in some cases involving US experts. Their reports are available on the ICES website[3] by searching on acronyms PGCCDBS, WKPICS, SGPIDS, WGRFS).

7. Provide guidance on key improvements in data or modeling approaches which should be considered when scheduling the next assessment.

The consensus Review Panel view is that for greater amberjack, the Stock Synthesis modelling framework still remains appropriate for the type of data available, and to allow flexibility to account for changes in size limits or IFQs that affect patterns of discarding in commercial and recreational fisheries. If it is agreed to continue with this approach, more work is needed to: i) more clearly express the relative quality of the different data inputs in relation to weighting, ii) to identify the minimum sufficient complexity of the model to provide robust advice on stock status, including identifying correlated parameters and applying fixed values, iii) to set appropriate priors to constrain model fits within bounds, and iv) to target work on improving the quality of the key data sets. Specific advice from the Review Panel for improving the SS3 implementation following the Review Panel meeting is to:

- Investigate correlations among model parameters (particularly for selectivity and retention), and either fix or provide informative priors for one at values that have some supportable evidence. If supportable evidence is unavailable and the parameter has a strong influence on the results, then a range of alternative fixed values should be investigated.

- For individual jitter starting points that resulted in different likelihood solutions, investigate which parameter estimates were affected that may also be fixed or provided with informative priors.

- Examine the CVs of parameter estimates. If the CV is large and the value has little influence on results, then choose a fixed value.

[3] http://www.ices.dk/publications/our-publications/Pages/Expert-Group-Reports.aspx

- Examine the time blocking of retention and selectivity for the fleets that converge at very high F values – consider very high to be values greater than 1.0 yr^{-1}, but preferably less than that. Consider adjusting the configuration of selectivity and retention of those fleets around the period of high F to see if the problem can be alleviated.

Improving the quality of length and age compositions for retained and discarded fish would help in fitting year class strength, selectivity and retention. However this is a challenge for greater amberjack due to the nature of the fisheries, and it is important to understand how improvements in precision of composition data translate into improvements in the quality of assessment outputs and advice, and the costs of sampling schemes that achieve this amount of sampling. Simulation modelling could be helpful in this regard.

Currently, the most influential relative abundance indices are from recreational and commercial fisheries, i.e. the same data sets used for estimating catch compositions and recreational catches, but filtered using information on species guilds in catches to try and identify trips where amberjack have a probability of being caught. Further work may be needed to identify potential biases in these approaches, for example where amberjack were initially targeted in a recreational trip but zero or low catch rates led to a switch to other areas or methods that do not catch amberjack. Other factors affecting catch rates in hook fisheries, particularly longlines (e.g. gear saturation, competition with other species) should be considered in evaluating if the commercial index series are reliable. Further investigation into the robustness of the design of the video surveys should also be carried out in relation to coverage of the stock.

Problems associated with poor identification of small amberjacks in observer programmes and self-reporting of discards by recreational fishers need to be addressed.

5. Conclusions and recommendations

For Gulf of Mexico Gag, the Review Panel has accepted the Stock Synthesis model as a basis for providing advice although has recommended the use of a fixed stock-recruit steepness value of 0.85, with the much higher Stock Synthesis estimate (as presented by the assessment team) and a lower value of 0.7 to indicate sensitivity of advice to uncertainty in this parameter. The perception of whether the stock of this protogynous hermaphrodite species is overfished is very sensitive to how SSB is calculated. The female-only SSB model indicates that the stock is no longer overfished in relation to any of the proposed reference points. However, the SSB-combined (male plus female) model indicates that the stock is overfished in relation to SSBSPR30%, but the SSB is marginally above SSBMSY. The stock does not appear to be undergoing overfishing. Recent management measures aimed at reducing the exploitation rate appear to have successfully lowered recent F values to below Fmsy and also Fspr30 for the agreed base model. The reduced IFQs in particular have caused a large increase in discard rate across the size range, and the benefits of the measure in part depend on survival rate of discards. However the estimated reduction in F, including the assumed discard mortality, is very large and unprecedented for the time series.

For Gulf of Mexico Greater Amberjack, Stock Synthesis is a more appropriate modeling framework than the previously-used ASPIC model for this stock, as it can estimate changes in selectivity and retention ogives which affect the composition of catches in a way that ASPIC cannot interpret. However, poor Stock Synthesis fits for many of the fishery composition data series, and unstable solutions, indicate model or data problems that could not be resolved during the Review Workshop. The Review Panel members are all of the opinion that Stock Synthesis remains an appropriate assessment approach for this stock, but the optimal configuration of the model had not been found at the time of the Review. As no base case assessment model could be identified, projections and stock status could not be evaluated. The Reviewers' have identified issues that should be addressed before the assessment model could be accepted as properly configured and consistent with standard practices, and consider that these could feasibly be addressed after the Review meeting to provide advice this year.

Appendix 1: Materials provided for review

DW and AW reports

SEDAR 33 SAR Section III	GAG Assessment Workshop final report	
SEDAR 33 SAR Section III	GAJ assessment Workshop final report	
SEDAR 33 SAR Section II	GAG Data Workshop report	
SEDAR 33 SAR Section II	GAJ Data Workshop report	

A large number of other background documents submitted to the DW and AW meetings were available on the SEDAR website.

Appendix 2: Statement of Work for Dr. Michael Armstrong (CEFAS)

External Independent Peer Review by the Center for Independent Experts

SEDAR 33 Gulf of Mexico Gag and Greater Amberjack Assessment Review

Scope of Work and CIE Process: The National Marine Fisheries Service's (NMFS) Office of Science and Technology coordinates and manages a contract providing external expertise through the Center for Independent Experts (CIE) to conduct independent peer reviews of NMFS scientific projects. The Statement of Work (SoW) described herein was established by the NMFS Project Contact and Contracting Officer's Representative (COR), and reviewed by CIE for compliance with their policy for providing independent expertise that can provide impartial and independent peer review without conflicts of interest. CIE reviewers are selected by the CIE Steering Committee and CIE Coordination Team to conduct the independent peer review of NMFS science in compliance the predetermined Terms of Reference (ToRs) of the peer review. Each CIE reviewer is contracted to deliver an independent peer review report to be approved by the CIE Steering Committee and the report is to be formatted with content requirements as specified in **Annex 1**. This SoW describes the work tasks and deliverables of the CIE reviewer for conducting an independent peer review of the following NMFS project. Further information on the CIE process can be obtained from www.ciereviews.org.

Project Description: SEDAR 33 will be a compilation of data, benchmark assessments of the stocks, and an assessment review conducted for Gulf of Mexico gag and greater amberjack. The review panel is ultimately responsible for ensuring that the best possible assessments are provided through the SEDAR process. The stocks assessed through SEDAR 33 are within the jurisdiction of the Gulf of Mexico Fishery Management Council and the state waters of Texas, Louisiana, Mississippi, Alabama, and Florida.. The Terms of Reference (ToRs) of the peer review are attached in **Annex 2**.

Requirements for CIE Reviewers: Three CIE reviewers shall have the necessary qualifications to complete an impartial and independent peer review in accordance with the tasks and ToRs described in the SoW herein. The CIE reviewers shall have expertise in stock assessment, statistics, fisheries science, and marine biology sufficient to complete the tasks of the scientific peer-review described herein. Each CIE reviewer's duties shall not exceed a maximum of 14 days to complete all work tasks of the peer review described herein.

Location of Peer Review: Each CIE reviewer shall participate and conduct an independent peer review during the SEDAR 33 panel review meeting scheduled in Miami, Florida during February 24-27, 2014.

Statement of Tasks: Each CIE reviewer shall complete the following tasks in accordance with the SoW and Schedule of Milestones and Deliverables herein.

Tasks prior to the meeting: The contractor shall independently select qualified reviewers that do not have conflicts of interest to conduct an independent scientific peer review in accordance with the tasks and ToRs within the SoW. Upon completion of the independent reviewer selection by the contractor's technical team, the contractor shall provide the reviewer information (full name, title, affiliation, country, address, email, and FAX number) to the contractor officer's representative (COR), who will forward this information to the NMFS Project Contact no later than the date specified in the Schedule of Milestones and Deliverables. The contractor shall be responsible for providing the SoW and stock assessment ToRs to each reviewer. The NMFS Project Contact will be responsible for providing the reviewers with the background documents, reports, foreign national security clearance, and other information concerning pertinent meeting arrangements. The NMFS Project Contact will also be responsible for providing the Chair a copy of the SoW in advance of the panel review meeting. Any changes to the SoW or ToRs must be made through the COR prior to the commencement of the peer review.

Foreign National Security Clearance: The reviewers shall participate during a panel review meeting at a government facility, and the NMFS Project Contact will be responsible for obtaining the Foreign National Security Clearance approval for the reviewers who are non-US citizens. For this reason, the reviewers shall provide by FAX (not by email) the requested information (e.g., first and last name, contact information, gender, birth date, passport number, country of passport, travel dates, country of citizenship, country of current residence, and home country) to the NMFS Project Contact for the purpose of their security clearance, and this information shall be submitted at least 30 days before the peer review in accordance with the NOAA Deemed Export Technology Control Program NAO 207-12 regulations available at the Deemed Exports NAO website: http://deemedexports.noaa.gov/.

Pre-review Background Documents: Approximately two weeks before the peer review, the NMFS Project Contact will send (by electronic mail or make available at an FTP site) to the COR the necessary background information and reports (i.e., working papers) for the reviewers to conduct the peer review, and the COR will forward these to the contractor. In the case where the documents need to be mailed, the NMFS Project Contact will consult with the COR on where to send documents. The reviewers are responsible only for the pre-review documents that are delivered to the contractor in accordance to the SoW scheduled deadlines specified herein. The reviewers shall read all documents deemed as necessary in preparation for the peer review.

Tasks during the panel review meeting: Each reviewer shall conduct the independent peer review in accordance with the SoW and stock assessment ToRs, and shall not serve in any other role unless specified herein. **Modifications to the SoW and ToRs shall not be made during the peer review, and any SoW or ToRs modifications prior to the peer review shall be approved by the COR and contractor.** Each reviewer shall actively participate in a professional and respectful manner as a member of the meeting review panel, and their peer review tasks shall be focused on the stock assessment ToRs as specified herein. The NMFS Project Contact will be responsible for any facility arrangements (e.g., conference room for panel review meetings or teleconference arrangements). The NMFS Project Contact will also be responsible for ensuring that the Chair understands the contractual role of the reviewers as specified herein. The contractor can contact the COR and NMFS Project Contact to confirm any peer review arrangements, including the meeting facility arrangements.

Tasks after the panel review meeting: Each reviewer shall prepare an independent peer review report, and the report shall be formatted as described in **Annex 1**. This report should explain whether each stock assessment ToR was or was not completed successfully during the SEDAR meeting. If any existing BRP or their proxies are considered inappropriate, each independent report shall include recommendations and justification for suitable alternatives. If such alternatives cannot be identified, then the report shall indicate that the existing BRPs are the best available at this time. Additional questions and pertinent information related to the assessment review addressed during the meetings that were not in the ToRs may be included in a separate section at the end of an independent peer review report.

<u>Contract Deliverables - Independent CIE Peer Review Reports</u>: Each CIE reviewer shall complete an independent peer review report in accordance with the SoW. Each CIE reviewer shall complete the

independent peer review according to required format and content as described in Annex 1. Each CIE reviewer shall complete the independent peer review addressing each ToR as described in Annex 2.

Specific Tasks for CIE Reviewers: The following chronological list of tasks shall be completed by each CIE reviewer in a timely manner as specified in the **Schedule of Milestones and Deliverables**.

1) Conduct necessary pre-review preparations, including the review of background material and reports provided by the NMFS Project Contact in advance of the peer review.
2) Participate during the panel review meeting at Miami, Florida from February 24-27, 2014
3) Conduct an independent peer review in accordance with the ToRs (**Annex 2**).
4) No later than March 14, 2013, each CIE reviewer shall submit an independent peer review report addressed to the "Center for Independent Experts," and sent to Mr. Manoj Shivlani, CIE Lead Coordinator, via email to shivlanim@bellsouth.net, and Dr. David Sampson, CIE Regional Coordinator, via email to david.sampson@oregonstate.edu. Each CIE report shall be written using the format and content requirements specified in Annex 1, and address each ToR in **Annex 2**.

Schedule of Milestones and Deliverables: CIE shall complete the tasks and deliverables described in this SoW in accordance with the following schedule.

24 January 2014	CIE sends reviewer contact information to the COR, who then sends this to the NMFS Project Contact
3 February 2014	NMFS Project Contact sends the stock assessment report and background documents to the CIE reviewers.
24-27 February 2014	Each reviewer shall conduct an independent peer review during the panel review meeting in Miami, Florida
14 March 2014	CIE reviewers submit draft CIE independent peer review reports to the CIE Lead Coordinator and CIE Regional Coordinator
28 March 2014	CIE submits CIE independent peer review reports to the COR
4 April 2014	The COR distributes the final CIE reports to the NMFS Project Contact and regional Center Director

Modifications to the Statement of Work: This 'Time and Materials' task order may require an update or modification due to possible changes to the terms of reference or schedule of milestones resulting from the fishery management decision process of the NOAA Leadership, Fishery Management Council, and Council's SSC advisory committee. A request to modify this SoW must be approved by the Contracting Officer at least 15 working days prior to making any permanent changes. The Contracting Officer will notify the COR within 10 working days after receipt of all required information of the decision on changes. The COR can approve changes to the milestone dates, list of pre-review documents, and ToRs within the SoW as long as the role and ability of the CIE reviewers to complete the deliverable in accordance with the SoW is not adversely impacted. The SoW and ToRs shall not be changed once the peer review has begun.

Acceptance of Deliverables: Upon review and acceptance of the CIE independent peer review reports by the CIE Lead Coordinator, Regional Coordinator, and Steering Committee, these reports shall be sent to the COR for final approval as contract deliverables based on compliance with the SoW and ToRs. As specified in the Schedule of Milestones and Deliverables, the CIE shall send via e-mail the contract deliverables (CIE independent peer review reports) to the COR (William Michaels, via William.Michaels@noaa.gov).

Applicable Performance Standards: The contract is successfully completed when the COR provides final approval of the contract deliverables. The acceptance of the contract deliverables shall be based on three performance standards:
(1) The CIE report shall completed with the format and content in accordance with **Annex 1**,
(2) The CIE report shall address each ToR as specified in **Annex 2**,
(3) The CIE reports shall be delivered in a timely manner as specified in the schedule of milestones and deliverables.

Distribution of Approved Deliverables: Upon acceptance by the COR, the CIE Lead Coordinator shall send via e-mail the final CIE reports in *.PDF format to the COR. The COR will distribute the CIE reports to the NMFS Project Contact and Center Director.

Support Personnel:

William Michaels, Program Manager, COR
NMFS Office of Science and Technology
1315 East West Hwy, SSMC3, F/ST4, Silver Spring, MD 20910
William.Michaels@noaa.gov Phone: 301-427-8155

Manoj Shivlani, CIE Lead Coordinator
Northern Taiga Ventures, Inc.
10600 SW 131st Court, Miami, FL 33186
shivlanim@bellsouth.net Phone: 305-383-4229

Roger W. Peretti, Executive Vice President
Northern Taiga Ventures, Inc. (NTVI)
22375 Broderick Drive, Suite 215, Sterling, VA 20166
RPerretti@ntvifederal.com Phone: 571-223-7717

Key Personnel:

NMFS Project Contact:

Ryan Rindone, SEDAR Coordinator
2203 N. Lois Avenue, Suite 1100
Tampa, Florida 33607
Ryan.Rindone@gulfcouncil.org Phone: 813-348-1630

Annex 1: Format and Contents of CIE Independent Peer Review Report

1. The CIE independent report shall be prefaced with an Executive Summary providing a concise summary of the findings and recommendations, and specify whether the science reviewed is the best scientific information available.

2. The main body of the reviewer report shall consist of a Background, Description of the Individual Reviewer's Role in the Review Activities, Summary of Findings for each ToR in which the weaknesses and strengths are described, and Conclusions and Recommendations in accordance with the ToRs. The CIE independent report shall be a stand-alone document for others to understand the weaknesses and strengths of the science reviewed. The CIE independent report shall be an independent peer review of each ToRs.

3. The reviewer report shall include the following appendices:

   Appendix 1: Bibliography of materials provided for review
   Appendix 2: A copy of the CIE Statement of Work

Annex 2: Tentative Terms of Reference for the Peer Review

SEDAR 33 Gulf of Mexico Gag and Greater Amberjack Assessment Review

1. Evaluate the data used in the assessment, addressing the following:
   e) Are data decisions made by the Assessment Workshop sound and robust?
   f) Are data uncertainties acknowledged, reported, and within normal or expected levels?
   g) Are data applied properly within the assessment model?
   h) Are input data series reliable and sufficient to support the assessment approach and findings?
2. Evaluate the methods used to assess the stock, taking into account the available data.
   a) Are methods scientifically sound and robust?
   b) Are assessment models configured properly and used consistent with standard practices?
   c) Are the methods appropriate for the available data?
3. Evaluate the assessment findings with respect to the following:

F) Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?

G) Is the stock overfished? What information helps you reach this conclusion?

H) Is the stock undergoing overfishing? What information helps you reach this conclusion?

I) Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?

J) Are the quantitative estimates of the status determination criteria for this stock reliable? If not, are there other indicators that may be used to inform managers about stock trends and conditions?

4. Evaluate the stock projections, addressing the following:

(d) Are the methods consistent with accepted practices and available data?

(e) Are the methods appropriate for the assessment model and outputs?

(f) Are the results informative and robust, and useful to support inferences of probable future conditions?

(g) Are key uncertainties acknowledged, discussed, and reflected in the projection results ?

5. Consider how uncertainties in the assessment, and their potential consequences, are addressed.

- Comment on the degree to which methods used to evaluate uncertainty reflect and capture the significant sources of uncertainty in the population, data sources, and assessment methods
- Ensure that the implications of uncertainty in technical conclusions are clearly stated.

6. Consider the research recommendations provided by the Assessment workshop and make any additional recommendations or prioritizations warranted.
   - Clearly denote research and monitoring that could improve the reliability of, and information provided by, future assessments.
   - Provide recommendations on possible ways to improve the SEDAR process.
7. Provide guidance on key improvements in data or modeling approaches which should be considered when scheduling the next assessment.

Annex 3: Tentative Agenda for

SEDAR 33 Gulf of Mexico Gag and Greater Amberjack Assessment Review
February 24-27, 2014
Miami, FL USA

Monday **[Note: Starting time revised to 9:00a.m.]**

1:00 p.m.	**Convene**	
1:00 – 1:30	**Introductions and Opening Remarks**	**Rindone**
	- Agenda Review, TOR, Task Assignments	
1:30 – 5:00	**Assessment Presentations and Discussions**	**SEFSC**
5:00 p.m. - 6:00 p.m.	**Panel Work Session**	**Powers**

Tuesday

8:00 a.m. – 11:30 a.m.	**Assessment Presentations and Discussions**	**SEFSC**
11:30 a.m. – 1:00 p.m.	**Lunch Break**	
1:00 p.m. – 3:30 p.m.	**Panel Discussion**	**Powers**
	- *Assessment Data & Methods*	
	- Identify additional analyses, sensitivities, corrections	
3:30 p.m. – 3:45 p.m.	**Break**	
3:45 p.m. – 5:00 p.m.	**Panel Discussion**	**Powers**
	- Continue deliberations	
	- Review additional analyses	
5:00 p.m. - 6:00 p.m.	**Panel Work Session**	**Powers**

Tuesday Goals**:** Initial presentations completed, sensitivities and modifications identified.

Wednesday

8:00 a.m. – 11:30 a.m.	**Panel Discussion**	**Powers**
	- Review additional analyses, sensitivities	
	- Consensus recommendations and comments	
11:30 a.m. – 1:00 p.m.	**Lunch Break**	
1:00 p.m. – 3:30 p.m.	**Panel Discussion**	**Powers**
	- Final sensitivities reviewed.	
	- Projections reviewed.	
3:30 p.m. – 3:45 p.m.	**Break**	
3:45 p.m. – 5:00 p.m.	**Panel Discussion/Work Session**	**Powers**
	- Review Consensus Reports	
5:00 p.m. - 6:00 p.m.	**Panel Work Session**	**Powers**

Wednesday Goals: Final sensitivities identified, preferred models selected, projection approaches approved, final results made available. Summary report drafts begun.

Thursday

8:00 a.m. – 12:00 p.m.	**Panel Work Session**	**Powers**

12:00 p.m. **ADJOURN**