Center for Independent Experts (CIE) Independent Peer Review of the 2012 Stock Assessment of Striped Marlin in the Western and Central North Pacific Ocean[1]

Tom Carruthers

t.carruthers@fisheries.ubc.ca

335 AERL 2202 Main Mall
University of British Columbia
Vancouver
Canada

December 2012

[1] Lee, H-H, Piner, K.R., Humphreys, R., Brodziak, J. 2012. Stock assessment of striped marlin in the western and central North Pacific Ocean. Billfish working group of the International Committee for tuna and tuna-like species in the North Pacific Ocean.

1. Executive Summary

The assessment document provides a clear description of a methodology with a number of central strengths. The modelling was undertaken using established software (SS3) that has been subject to extensive review in stock assessments elsewhere. Where available, the assessment makes use of new information on the biology and ecology of the striped marlin stock. The report provides an excellent introduction to the fishery and presents a complete account of the different stock structure and mixing hypotheses. The assessment modelling itself is very detailed, particularly in the attempt to capture the historical fishing mortality at size of multiple fleets. In general, the authors do an excellent job of highlighting and discussing many of the implicit assumptions of the model. Such an assessment can produce a large quantity of output. However, the results of the assessment are clearly presented in a range of informative figures and tables including several that allow the reader to easily make cross-stock comparisons. A comprehensive range of sensitivity analyses and projections is another key strength of the assessment.

There are however important areas in which greater detail is required in order to have confidence over the assessment results. In particular, much greater transparency is needed in the description of CPUE standardization approaches used to construct relative abundance indices in order to understand whether these were calculated correctly. The account of the size-frequency data is also too cursory for a reader to understand whether these data are sufficiently informative to support a large and intricate component of the assessment model.

There is not a clear justification for assessment model structure such as seasonality and the inclusion of regional fisheries. In some cases the base-case assessment model appears to include complexity that is demonstrated to be unnecessary by sensitivity analysis. In general the model seems very complex in its approximation of fishery dynamics relative to the simplicity of the spatial population assumptions. It is problematic therefore that the document does not provide sufficient assurance that the model is not overparameterized; that it converges reliably and can robustly estimate management reference points. Let us assume that the purpose of stock assessment is the provision of reliable management advice. It is difficult to know whether the assessment model adopted here can be expected to perform better than more simple approaches.

Various parts of the methodology contain disparities in assumptions. For example, some important standardization models assume a viscous stock with regional abundance trends and seasonally variable biomass that are not fully accounted for by the assessment model.

There are also some important overarching problems. The lack of clear management objectives for the striped marlin stock prevents assessment results from being presented in a meaningful framework and may hinder the development of quantitative tools to support decision making.

2. Background to the review

Striped marlin (*Tetrapturus audax*) is one of six species of billfishes commonly harvested multi-nationally from commercial and recreational fisheries in the western and central Pacific Ocean regions. Fishery management requires high quality science to effectively manage and conserve our living marine resources, and the scientific peer-review of stock assessments by external Center for Independent Experts (CIE) expertise is an important process in the determination of best scientific information available (from the Statement of Work Appendix 2).

Important fisheries for striped marlin have operated in Western and Central North Pacific Ocean (WCNPO) since the early 1950s. Catches in the period before 1975 are considered to be in the range of 4000-8000mt (the 'equilibrium catch' prior to 1975 is assumed by this assessment model to be 5000mt). Since 1975, catches in the longline fishery fluctuated but the trend remained broadly flat until the early 1990s. The substantial contribution of the driftnet fishery after 1970 (approximately 1/3 of catch) lead to peak catches around 10,000mt in the mid 1980's. After 1992, catches in all fleets can be observed to decline to current (2010) total levels of around 2600mt. While simplistic, the consistent decline in both catches and catch rates of key fleets (*e.g*. Japanese longline fleet, Kanaiwa *et al*. 2011) provides some reason for concern about the status of the WCNPO striped marlin stock.

Striped marlin is a highly migratory species occupying a large spatial range over which large discrepancies in population density and size structure are likely to occur. While spatial structure is a central challenge in the approximation of population dynamics, size-dependent exploitation by multiple fleets poses difficulties when attempting to represent fishing dynamics. An additional challenge for stock assessment is the paucity of reliable data prior to 1975, a period over which a relatively large degree of stock depletion may have occurred.

The previous assessment in 2007 (MAR&SWO 2007; Piner *et al*. 2006, 2007) was the first detailed attempt to characterize fishing /population dynamics and estimate current stock status. The central conclusion of the 2007 assessment was that spawning potential ratio (a measure of 'health' of a fish stock) was less than 10% that of unfished levels. It proved difficult to interpret these results with confidence due to the generally sparse biological and ecological information available to support the assessment. Central uncertainties included "stock structure, spawner-recruit resilience (h), natural mortality (M) and the growth rate of the species" (from the document under review here, Lee *et al*. 2012, hereafter referred to as 'the document' or 'the report'). Subsequently a number of studies have provided more credible estimates of these inputs (*e.g*. Sun *et al.* 2011a,b,c; Piner and Lee 2011a,b) potentially improving the basis for a revised stock assessment.

The current assessment considerably updates and expands on the 2007 analysis taking into account new information on life-history, recent total catch, catch composition and relative abundance indices. The current assessment prescribes a higher level of stock productivity with higher base-case assumptions regarding natural mortality rate and steepness (age specific natural mortality ranging from 0.54-0.38 as opposed to 0.3 across all ages in the 2007 assessment; steepness of 0.85 as opposed to 0.7 in 2007).

3. Description of the Individuals Reviewers Role in the Review Activities

A detailed description of the reviewer's role can be found in the Statement of Work (Appendix 2). The supporting documents, many of which are listed in the bibliography (Appendix 1) were received on the 26th October 2012. The main assessment report was received on the 10th of November 2012. Most of the remaining documents listed in the Bibliography can be found on the ISC website.

4. Summary of findings in regard to TORs (weaknesses and strengths)

4.1. Review of the assessment methods: determine if they are reliable, properly applied, and adequate and appropriate for the species, fisheries, and available data.

A specific discussion of these issues is included in the detailed break-down of Section 4.2 below. It is hard to judge whether the data available are of sufficient quality to support an assessment of this complexity. Given the lack of information regarding model convergence it is difficult to know whether the assessment methods are reliable, properly applied and appropriate for the species (Section 4.2.1), fisheries and the available data (4.2.2).

4.2. Evaluate the assessment model configuration, assumptions, and input data and parameters (fishery, life history, and spawner recruit relationships): determine if data are properly used, input parameters seem reasonable, models are appropriately configured, assumptions are reasonably satisfied, and primary sources of uncertainty accounted for.

4.2.1 Model configuration and assumptions

Description of the model

The authors provide a comprehensive description of the model in the main text of the document but the transition equations (page 31) do not include the seasonal structure. This raises a number of questions about model assumptions. These are probably ignorable but it would be instructive to know when recruitment was assumed to have occurred during a modelled year.

A seasonal population dynamics model

From the assessment document it is hard to completely understand the rationale for a seasonally disaggregated assessment. It is argued that a seasonal model is necessary to (1) reliably estimate selectivities and (2) account for difference in the magnitude of catch between seasons (page 21, paragraphs 3 and 4). That size composition of catches differs between seasons is not necessarily a problem for an annual assessment. The definition of selectivity simply changes to the fraction of modal fishing mortality rate exhibited on an age class over the course of a year. That the magnitude of catch differs between seasons is also not necessarily a problem for an annual assessment; perhaps fishing effort is greater in certain seasons. The justification for a seasonal model is usually based on the seasonal interaction between the stock and the fleet among years. If the seasonality in the stock is constant over time (roughly the same seasonal pattern in abundance and size composition among years) and the distribution of fishing is also constant seasonally over time (roughly the same fraction of effort in each season among years) then an annual assessment may not be strongly biased. If however there are large fluctuations in the seasonal pattern of fishing (between years a very different sub population structure is being fished) then it may be important to model seasonal population dynamics. However this rationale is not referred to by the authors and there is insufficient information in the assessment document to understand whether the seasonality in fishing and/or the population are changing dramatically over time. It would be interesting to see the difference in estimated reference points from an annual model with the other assumptions held constant.

Disparity in the magnitude of spatial assumptions and mortality-at-size assumptions

I agree with the authors' decision to fix a number of inputs that are generally not well estimated such as steepness, mortality at age and maturity-at-age. It is interesting that projections (and therefore the basis for management advice) depend most heavily on a very small number of estimated parameters, in particular stock size (*R0*) and recent recruitment deviations (that may not be that well estimated). The remaining model complexity (napkin arithmetic points to over 120 parameters) is spent approximating size-specific fishing dynamics. It is possible that this demands too much of the size composition data and it seems disproportionately concerned with approximating size-specific mortality, particularly considering that the model relies on relatively large spatial assumptions.

A general comment on model complexity

From my perspective, the tendency for *ad-hoc* adjustment of model structure to remove fine-scale residual patterns may be problematic and inconsistent with the broader objectives of stock assessment: "Seventeen fisheries were initially defined but further analysis indicated that a residual pattern and quarterly size observations from the Japan other fishery showed a substantial seasonal pattern of larger fish caught in the first two seasons (see Section 3.5 below on length frequency data and Figure 5). Seasonality in selectivity was modeled by splitting the Japan "other fishery" into two seasonal fisheries corresponding to seasons 1-2 and 3-4 of the calendar year in order to reduce the influence of the misfit" (Page 20, last paragraph).

It may be possible to tweak assessment models to remove patterns in residual errors. But the objective of stock assessment is not usually to provide a complete account of all of the historical complexities of the system. Generally we wish to robustly capture stock size and productivity in order to make reliable decisions about how to manage the stock. Adding parameters to chase down areas of model misfit will certainly increase the challenge for the numerical optimization, may lead to over-fitting and spurious predictions without necessarily improving the reliability of decision making. While I acknowledge the comprehensive attempt made by the stock assessment scientists to account for important changes in fishery dynamics in terms of residual errors, I am concerned that there may be little benefit in terms of the provision of reliable management advice. Other areas where complexity has been added to remove residual patterns is the blocking of historical selectivities. My experience reviewing similar models applied to stocks elsewhere (with much fewer fleets) has been that the model may be highly overparameterized. There is an onus in such instances on providing conclusive evidence of convergence that in my view is not reflected in the stock assessment document.

In Section 5.1 of the results (page 35) it is stated that: "There is no evidence of substantial differences in the scaling parameter (R0) and total likelihood showing a better fit (Figure 12). Based on these results, the BILLWG concluded that the base-case assessment model is relatively stable with no evidence of lack of convergence to the global minimum". In my view this does not constitute a sufficiently detailed assurance of convergence. In the most overparameterized models I have reviewed previously, the scale of the population represented by *R0* remains consistently estimated under jittered runs. The fact that the objective function does not change may also mask a serious problem (*e.g*. a very flat and poorly-defined objective surface). It would be much more informative to see the variance in the estimates of quantities of management reference points (*e.g*. *MSY*, $B_{current}/B_{MSY}$ and $F_{current}/F_{MSY}$) from different starting values. Other important diagnostics of model overparameterization that could be included here are parameters estimated at their bounds, strong posterior cross-correlation among parameters, poor MCMC mixing (high auto-correlation requiring heavy thinning to satisfy convergence diagnostics) unrealistically high precision in model predictions, high sensitivity in model predictions to credible changes in input parameters and an inability to recover known parameter values from simulated data (a built-in feature of SS3). To have confidence over the model predictions it would be highly informative to have a more complete account of whether the model suffers from such phenomena.

In several instances, reducing the complexity of the model or the way in which data are included, does not significantly alter results (*e.g*. insensitivity to spatially disaggregated index for the Japanese longline fleet; section 5.6.1.1, page 39). It is not clear why the base-case model should be the more complicated of the two configurations that produce similar results.

The objective function

The authors make a sensible decision to reweight the effective sample size of the length composition data to reflect non-independence in observations. This helps to combat a well-established problem in which the analysis becomes dominated by the composition data.

The authors state that there are three components to the likelihood function: the total catch data, CPUE indices and the length-frequency data. I suspect that this is not strictly correct and that there are a number of other likelihood components that go unmentioned. For example, penalties for parameters as they reach their bounds or as predicted stock size approaches zero. The difficulty for most assessments undertaken in SS3 to completely describe the objective function remains a central criticism I have of the software.

4.2.2 Input data

Length-frequency data

Due to the relatively complex fishery dynamics (multiple fleets, multiple seasons, blocked selectivities *etc*.) the model may be strongly dependent on the quality of the length-frequency data. Section 3.5 (page 22) provides a brief description but it is not easy to gain an intuition about the amount of data gathered and the process used to derive the lengths. Additional information would be useful here to allow a reader to better understand to what extent these data can be expected to inform the more intricate aspects of the assessment model.

The division of relative abundance indices into temporal blocks

It is argued that the Japanese longline indices should be divided into three periods (1975-1986, 1987-1999 and 2000-2009) to “account for changes in operation, hook-per-basket (HPB) distribution, targeted fish and length distribution of catch”. It is usually the role of standardization to remove these confounding influences. Indeed Kanaiwa *et al*. (2011) account for HPB in their standardization. By severing the abundance time series into three parts, long-term information regarding depletion is largely removed and three nuisance parameters (*q*'s by time period) must be estimated instead of one. The likely result is that depletion is now informed to a much greater extent by other data that may not be as reliable. For example, the inference regarding total mortality rate *Z,* from in the catch composition data.

The inclusion of relative abundance indices for multiple fleets

There is only one trend in real abundance and several standardized indices that each provide a different inference of stock trend. Let us assume that the objective of the standardizations is to produce an index of population-wide abundance. Assuming that the stock is fully mixed and vulnerability schedules are comparable among fleets there are two possible conclusions: that (1) all but one of the standardization methods are not operating correctly or (2) all of the standardization methods are not operating correctly. In such cases it is not defensible to fit the model to multiple sets of derived data of which the majority are known to be incorrect. The assessment report includes a note to this effect (page 28, last paragraph) but then continues with a multiple index approach.

On a practical level, the objective surface on which the optimization algorithm operates usually becomes less well-defined with multiple local minima at different parameter vectors that suit particular abundance indices.

On a theoretical level, the inclusion of multiple abundance indices may increase the strength of the spatial assumptions of the assessment. Clearly a spatially aggregated model is to some extent inconsistent with known spatial characteristics such as size structuring, regional abundance trends and spatial heterogeneity in abundance. In particular spatial heterogeneity in abundance is evident from spatial plots of CPUE among fleets. The implicit assumption of the spatially aggregated model is that while population density may vary in space, the distribution of the fleet is constant in relation to this regional abundance (effort is distributed consistently on the population). Fleets with the most complete spatial coverage sample a greater range of the stock and indices are weighted in proportion to regional abundance (in a GLM with marginal time and area effects for example). Including an index for a regional fishery increases the spatial assumptions of the assessment: we now have to assume that regional trends reflect overall population trends. Additionally the inclusion of a regional index may provide extra weight to a particular area of relatively low abundance and may bias the assessment by the location of observations. For this reason those indices that are applicable to small sub-areas such as the Japanese coastal large-mesh drift fishery (Yokawa and Kimoto 2011) should not be used to infer-population wide abundance trends in a base-case assessment. This is particularly the case if the density of the population in these areas varies among years due to temporal changes in rates of migration (as implied by the year x season interaction effects modelled by Yokawa and Kimoto 2011).

Similarly, the theoretical basis for including three regional abundances indices for the Japanese longline offshore fishery (JPN_DWLL1, JPN_DWLL2, JPN_DWLL3) is not clear. It is theoretically inconsistent to incorporate such data in an assessment model that assumes that there are no regional abundance trends. However, it may be argued that while regional trends in abundance exist we may approximate stock dynamics by a single area model. Since these regional indices are essentially time x area interactions it would be better to undertake a single CPUE standardization including these interactions. This would derive a single index weighted by the predicted level of abundance in each region. Currently, the assessment assigns equal prior weight to these indices when in fact one may be tracking the abundance of a substantially smaller fraction of the population.

It is indicated in the assessment report (page 28, last paragraph) that the reason for including multiple series was due to a lack of consensus in expert judgement regarding the representativeness of the different indices. An attempt was made to evaluate the different indices objectively based on correlation analysis and sensitivity analysis. The rationale for why these approaches should be helpful was not provided. It might be better to undertake a basic simulation exercise. In their simulation analysis of CPUE for tuna and billfish fisheries in the Atlantic (Carruthers *et al*. 2010), there were clear grounds to favour some fleets over others due to their relatively complete spatial coverage and recording of catch rate covariates over time. For example, if it can be assumed that the population is spatially well mixed, CPUE standardization methods with marginal area effects were best informed by the Japanese longline fleet that has the most complete spatial coverage over time and has relatively complete records of hooks-per-basket.

Japanese longline CPUE 1975-2010, Kanaiwa *et al*. 2011.

Kanaiwa *et al*. (2011) provide a cursory account of their standardization method and provide insufficient information to reproduce the method. Fundamental equations are missing. For example, it is not clear how annual relative abundance and standard error were calculated. Future standardization approaches should include the equation $I_t = \ldots$ where I is the derived index and t the time subscript. Kanaiwa *et al*. (2011) provide no information about how they dealt with potentially critical problems regarding the interpretation of catch rate data. Did they account for differences in the size of areas? How were the confounded year x area and quarter x area interactions dealt with in the calculation of annual index? Did the authors account for different sample sizes among strata (*e.g*. Campbell 2006)? It is not clear why the model should use numbers/effort as opposed to weight/effort as a unit of CPUE since (as far as I can tell) the indices are compared with biomass in the assessment model.

It is not clear why the standardization was undertaken on a spatially aggregated dataset as opposed to the more detailed trip-level data that contains 1x1 degree detail and hooks-per-basket covariate information. I suspect this was due to the computational constraints of modelling time x area interaction effects. If aggregating data to 5x5 degree cells was necessary to model interactions it would be desirable to understand whether this additional model complexity was necessary. Unfortunately Kanaiwa *et al*. (2011) provide no plots of the time x area interactions (or in fact any plots of indices) to examine whether there are differences among the predicted regional abundance trends. It should be noted that modelling CPUE with time x area interactions is based on spatial assumptions that are inconsistent with that of the assessment model; namely that there the stock is not well mixed and there are regional abundance trends in contrast to the assessment model that assumes a perfectly mixed stock.

The absence of basic figures is inexplicable. In a paper generating relative abundance indices it is strange not to include a graph of the derived abundance trends over time. Such plots should include a representation of uncertainty and the sensitivity of the derived index to credible alternative assumptions.

It is my experience that accounting for the marked changes in set depth of fleets such as Japanese longliners has a profound impact on the standardization of billfish CPUE. As noted in this document (Page 18, paragraph 1) longlining in the 1970s shifted towards deeper sets to target bigeye, a phenomenon also observed in longline fishing in the Indian and Atlantic Oceans (Miyake 2004). The expected effect is a marked increase in the inferred relative abundance for species inhabiting shallower waters. It is therefore surprising that the standardization of Kanaiwa *et al*. (2011) that accounts for both spatial effects and depth effects leads to an inferred relative abundance that is so similar to the unprocessed nominal CPUE (see Figure 1 below). Without providing CPUE data and a reproducible method it is not possible to investigate this atypical result and I remain concerned about this important input to the stock assessment.

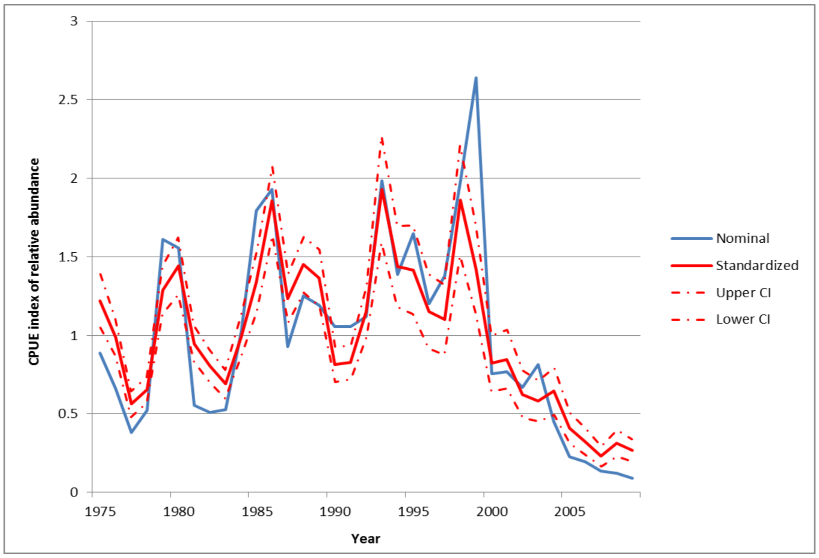

Figure 1. Base-case standardization of CPUE versus nominal CPUE for striped marlin in the WCNPO presented by Kanaiwa *et al*. (2011). The dotted lines represent the 95% confidence interval assuming that the standard deviation reported in Kanaiwa *et al*. (2011) refers to the mean and can be interpreted as a standard error.

The reliance on AIC to select the structure of the GLM standardization model is problematic. In general AIC is known to lead to the selection of GLM models that are overparameterized (Kadane and Lazar 2004). This tendency is magnified in the specific case of CPUE standardization because the objective is the extraction of a reliable relative abundance index not the prediction of the next CPUE observation (Carruthers *et al*. 2010); AIC is suited to the latter. Instead of model selection criteria it is desirable to use current ecological and fishery knowledge to select a defensible model (or small set of competing models) *a priori* rather than use AIC to undertake an ad-hoc search for GLM model that selects the best predictor of CPUE observations. For example: (1) we may have reason to believe that striped marlin inhabit different depths in different areas due to spatial changes in the depth of the mixed layer (supporting the inclusion of depth x area interactions); (2) tagging studies may indicate that the stock is viscous and spatial differences in the intensity of fishing would imply that regional abundance trends could be different (supporting the inclusion of time x area interactions).

Japanese coastal large-mesh drift fishery CPUE, Yokawa and Kimoto 2011

Similarly to Kanaiwa *et al*. (2011) above, Yokawa and Kimoto (2011) do not present basic calculations that are necessary to evaluate their standardization approach. There is no clear explanation of how they calculated their index. In particular it would be useful to know how they derived an annual index given a year effect that is confounded with the year x quarter interaction effects. This is not an ignorable matter and entirely determines the credibility of the index. In many standardization papers that I have previously reviewed, the marginal year effects are interpreted directly without consideration of the confounded interactions leading to spurious abundance trends.

It is not clear why the categorical effects were estimated in log-space. This is unusual and it would be useful to see the SAS code for this modelling.

The authors provide no explanation for the modelling of a year x quarter interaction. It implies that the stock is seasonally transitory in this fishing area over years. It follows that Yokawa and Kimoto (2011) acknowledge that the stock moves in and out of this area seasonally and that a low/high annual abundance could be driven by migration. This is problematic because it requires a level of spatial detail absent in the spatially aggregated assessment model. In my view these assumptions are sufficiently disparate to preclude the use of this index in the assessment.

Japanese high sea large-mesh driftnet CPUE, BILLWG 2011b

The 2011 Report of the Billfish Working Group Workshop (BILLWG 2011b) is provided in the stock assessment as a reference for the Japanese high sea large-mesh driftnet fishery. This is an update of the standardization of Yokawa (2005) that derived an index from 1977-1993. Yokawa (2005) provides more information regarding the underlying data. For example it is noted that a moratorium of large scale drift net fishing in the open ocean in 1993 led to all but 4% of catches being taken in a single near shore area. Unfortunately no information is provided about how this issue was dealt with in either the analysis of Yokawa (2005) or the update (BILLWG 2011b). Similarly to the standardization papers above it is not clear exactly how the index was calculated or the sensitivity of the index to other credible assumptions about stock and fishery dynamics. Again, it is not clear why the categorical effects were estimated in log space.

Taiwanese distant-water longline fishery 1967-2009, Sun *et al.* 2011d

The lack of transparency in the description of standardization methods described above is also applicable here. Similarly, no sensitivity analyses are conducted. There are additional curiosities however. The equation describing the GLM standardization model does not include a transformation of the nominal catch rate data which implies that the authors were applying a normal error model to a variable that cannot be negative (later they refer to a log-normal error distribution in contradiction to their GLM equation). The GLM model applied includes latitudinal and longitudinal marginal effects and a latitude x longitude interaction effect. They do not state whether these are continuous variables or discrete categorical spatial blocks. If they are categorical there is no reason to model the marginal effects as they are already accounting for areas defined in two-dimensional geographic space. We do not know how the authors accounted for potential sources of bias such as uneven spatial distribution of observations and differences in sizes of modelled areas.

The authors do not mention whether covariate set depth data are also available for these fisheries. Given the spatial range of this fishery, if depth and other covariate information regarding gear (*e.g.* bait) were available, this CPUE series could offer an interesting and credible alternative to the Japanese longline index as a primary source of information for stock-wide abundance trends.

Hawaiian Pelagic longline observer CPUE 1995-2009, Walsh and Lee 2011.

Of the indices used in the assessment Walsh and Lee (2011) provide the most comprehensive account of their standardization approach. Given that they did not model year interaction effects it is easier to deduce how they may have calculated their index (however the reader still does not know how they dealt with unbalanced sampling and areas of different size). The analysis provides a number of valuable findings in particular the relative impact of sea surface temperature and set depth.

In regional index standardization of this type it would be useful to have illustrations of the spatial extent of the observations in order to assess how general indices are likely to be in regard to population-wide abundance. Additionally several of their residual plots show distinct structuring (clumping of plotted points) that suggest population or fishery characteristics not accounted for by the GLM modelling (*e.g*. the plot of sea surface temperature observed vs. standardized residuals).

Walsh and Lee (2011) conclude that "…striped marlin catches and catch rates have decreased considerably in the last 15 years. The relatively similar estimates of change in catch per set and CPUE and the finding that the standardized trends were highly correlated with differences that could be ascribed to operation changes in this fishery reinforces this conclusion". This is a strange statement in the context of CPUE standardization. If the indices are correlated with operational trends then the apparent changes may be an aberration of observation processes (that we wish to standardize for) rather than population changes. For example, consider a fishery in which a species is increasingly targeted at a particular depth. This operational shift may lead to apparent declines in species inhabiting other depth ranges. In the case of either species the correlation does not provide corroboration of the apparent trend in CPUE.

4.2.3 Input parameters

The fundamental biological and ecological relationships assumed by the model appear to be sound and the supporting documentation is generally detailed and carefully presented.

Growth, maturity and the stock-recruitment relationship

The assessment is well supported by the growth studies of Sun *et al.* (2011b; 2011c), the spawning analysis of Sun *et al*. (2011a) and the stock recruitment work of Brodziak *et al*. (2011). The estimate of steepness 0.87 may be a good base-case assumption but the '+/-' 0.05 appears to be quite a precise range for a quantity that is generally poorly informed. This is recognised in the sensitivity analysis and the range of values considered is much wider.

Natural mortality rate

Piner and Lee (2011a;b) provide a comprehensive account of the derivation of their natural mortality rate-at-age estimates. Their method is interesting and defensible relative to other papers on the subject and they acknowledge potential shortcomings of the approach. As they note, the same approach has been adopted elsewhere and has the core benefit that it provides a measure of uncertainty with which to bracket sensitivity analyses.

Sensitivity analyses

In general, a very comprehensive range of sensitivity analyses are included in the assessment document and these help the reader to understand more about the behaviour of the model.

In some cases changes in input parameters are used as the basis for sensitivity analysis. For example natural mortality rate: "Results indicate that models for both natural mortality rates fit worse by a moderate amount for length compositions (9 and 4 likelihood units worse than base case for high M and low M, respectively). However, fit to CPUE series appeared similar based on the small changes in likelihood (< 2 likelihood units). In summary, total likelihood favors the base case model". Various structural aspects of the base-case model have been changed *ad-hoc* in order to reduce residual error given base-case inputs (such as natural mortality): "The authors note that many additional sensitivity runs were conducted in the development of the base case (*e.g*. bin definitions, initial conditions, alternative data sets *etc*.) that are beyond the scope of this paper to describe" (page 30, paragraph 3). It is therefore not surprising that alternative inputs lead to poorer fit and it may not be possible to interpret this as evidence with which to favour one natural mortality rate assumption over another.

4.2.4 Accounting for uncertainty

In general the assessment does a good job of expressing uncertainty by providing confidence intervals for predicted quantities and undertaking a range of sensitivity analyses. As noted in the report, structural uncertainty is not well accounted for and it would have been instructive to see the difference in predictions of key reference points generated from a more simple assessment model (*e.g.* with/without seasonal structure).

The assessment document refers to Bayesian methods of characterising uncertainty: "The structure of the model allows for Bayesian estimation processes and full integration across parameter space using the Monte Carlo Markov Chain (MCMC) algorithm" (page 23, paragraph 1). However this is not mentioned elsewhere in the document with the exception of the 'MCeval' lines of the SS3 starter file (page 104). There is no formal reference to chain thinning, convergence diagnostics or other matters that are necessary when interpreting MCMC outputs. A number of other comments imply that the MLE estimates and a hessian approximation to the standard error were used to construct the estimates of uncertainty expressed in tables and figures. Additionally the use of the term 'confidence interval' (a frequentist concept) implies that the MCMC run was not used to provide estimates of uncertainty.

It would be desirable to have the Bayesian posterior estimates since they are much more straightforward to understand and interpret. It may also be the case that the MLE estimate (the posterior mode) is not a suitable estimate of an expected value if the posterior is strongly skewed (in which case a posterior median is preferable). Several Bayesian outputs would also be useful in diagnosing model overparameterization, in particular convergence diagnostics and the joint posterior parameter cross-correlation plots. However it is likely that given the complexity of the model the MCMC evaluation would take several days (perhaps weeks) to run in order to satisfy convergence diagnostics and perhaps this is reason why these results were not included.

4.3. Comment on the proposed population benchmarks and management parameters (e.g., MSY, Fmsy, Bmsy, MSST, MFMT); if necessary, recommended values for alternative management benchmarks (or appropriate proxies) and clear statements of stock status.

The stock assessment uses standard metrics such as spawners per recruit and spawning stock biomass and fishing mortality rate relative to *MSY* levels. These metrics are widely applied elsewhere and are relatively easy to interpret. Standard outputs such as Kobe plots (*e.g*. Figures D and E, page 7) are included here that provide a transparent account of predicted historical stock status and exploitation level.

It appears that there are no agreed management reference points for this stock: "No target or limit reference points have been established for the WCNPO striped marlin stock under the auspices of the

WCPFC" (page 45). However the authors make sure to use standard definitions of over/underfishing and over/underfished to describe the outputs of the assessment. In future assessments it would be highly desirable to have not only target and limit reference points but also acceptable probabilities for exceeding these levels. These definitions would allow assessments to phrase results in a clear way in direct reference to management objectives (for example probability of an MCMC projection not exceeding the limit reference point in a projected year). Other related benefits include the foundation for undertaking value-of-information analysis and management strategy evaluation.

4.4. Evaluate the adequacy, appropriateness, and application of the methods used to project future population status.

The assessment report includes a detailed series of projections including both input and output control scenarios. The decision to limit projections to a relatively short time period of 5 years is defensible given the high level of recruitment variability and lack of apparent autocorrelation in recruitment. The suite of SS3 tools for conducting projections has been subject to review and revision in many other stock assessment settings and can certainly be considered to be adequate and appropriate.

I am however concerned that projections may be optimistic and do not express a degree of uncertainty that is intuitive given the historical data and historical predictions of the assessment model. Over the years 2005-2010, the base-case model has predicted consistent declines in spawning stock biomass from catches as low as 2560t (Table 1 below, Table 4 of the report). Recent recruitment is estimated to be high which presumably drives the model projections that lead to the conclusion "When catch is reduced 20% from current level (average 2007-2009) which is about 2,500mt, the stock is projected to have zero chance to fall below 2012 level for both states of nature" (page 43). To have zero probability of falling below 2012 levels by 2017 implies a degree of certainty in projections at 2,500mt y^{-1} that does not appear to be consistent with the past data and projections (Table 1 below, Table 4 of the report). In my view this indicates that the projections may be somewhat optimistic and that recovery of the stock at 2,500mt may not be as pronounced as predicted here.

Table 1 (Table A of the report, page 4). Observed catch and base-case model predictions of stock status.

Year	2004	2005	2006	2007	2008	2009	2010	Mean[1]	Min[1]	Max[1]
Reported Catch	4047	3703	3706	3195	3691	2560	2560[2]	6011	2560	10528
Population Biomass	11679	9545	10371	8430	7414	5335	6625	14141	5335	24886
Spawning Biomass	1731	2010	1992	1824	1625	1106	938	2439	909	5104
Relative Spawning Biomass	0.64	0.74	0.73	0.67	0.60	0.41	0.35	0.90	0.33	1.88
Recruitment (age 0)	116	434	125	204	133	349	326	453	116	1620
Fishing Mortality	0.58	0.56	0.62	0.58	0.86	0.84	0.75	0.79	0.53	1.46
Relative Fishing Mortality	1.22	0.95	0.92	1.01	0.95	1.41	1.37	1.30	0.86	2.38
Exploitation Rate	35%	39%	36%	38%	50%	48%	38%	44%	29%	69%
Spawning Potential Ratio	19%	19%	17%	19%	12%	13%	14%	14%	7%	21%

4.5. Suggest research priorities to improve our understanding of essential population and fishery dynamics necessary to formulate best management practices.

Increasing the transparency of the CPUE standardization process may help to arrive at methods that best characterize declines in relative abundance. This could involve making CPUE data available or allowing standardization models to be investigated interactively during a stock assessment meeting.

A research priority is undertaking simulation evaluation to identify management strategies and assessment models that are robust to uncertainty in stock and fishery dynamics. This type of simulation may also provide a formal basis for selecting abundance indices and understanding the value of additional information in terms of making reliable management decisions. For example, how cost effective is a conventional tagging programme (given the 1% return rates) in improving management performance?

It may be useful to investigate the cost-efficacy of newer tagging technologies that do not rely on reporting of tags (*e.g*. an integrated PIT tag – reader system) that may offer an alternative source of information regarding current stock size and fishing mortality rate.

5. Conclusions

Strengths

The assessment was conducted using established software that has been subject to extensive review in other stock assessment settings.

The assessment makes good use of updated information regarding the biology and ecology of the stock.

The assessment is very detailed particularly in its attempt to capture the historical fishing mortality at size of multiple fleets.

The results of the assessment are clearly presented in a range of informative figures and tables.

In general, the authors do an excellent job of describing the model and discussing the implicit assumptions.

A comprehensive range of sensitivity analyses and projections is another key strength of the report.

Weaknesses

In many cases an assessment may be evaluated in terms of its ability to predict the trends inferred by the relative abundance indices that are provided by stakeholders. These derived data may therefore directly determine one of the most important reference points of the assessment, stock depletion. Small changes in the standardization approach can have dramatic impacts on the inferred abundance trends. In most cases the standardization research used by this report is not presented in a manner that is consistent with basic standards of scientific publication elsewhere. Methods are not described in sufficient detail to be reproducible. The inherent assumptions of the standardization methods are not made clear or discussed. It is not possible to understand whether the authors of these indices have satisfactorily accounted for a number of potentially serious biases that affect CPUE standardisation that could preclude the use of the index in stock assessment.

The gathering and processing of critical inputs such as length composition data are not described in sufficient detail for a reader to gain an intuition of the quality of these data.

The complexity of the modelled fishing dynamics appears disproportionately high relative to the probable quality of the data and the size of the assumptions regarding spatial population dynamics. Given their experience and expertise there is good reason to trust the authors' judgement. However in its own right the document does not include sufficient detail regarding convergence diagnostics and evidence against model overparameterization. These are a precondition for interpreting assessment outputs and based on the report alone it is difficult to be confident about the robustness of the assessment results.

The assessment document does not include enough detail about model structure such as seasonality (*e.g*. seasonal transition equations that include recruitment).

In several cases the base-case model appears to include unnecessary complexity which is undesirable (lack of sensitivity to spatial disaggregation of the Japanese longline index for example).

The use of regional data to infer population-wide stock dynamics is questionable and may lead to results that are biased by regional characteristics.

The CPUE standardization methods of different fleets rely on assumptions about population dynamics that conflict with one another and the assessment.

The index derived from the fleet with probably the best spatio-temporal coverage (the Japanese longline fleet) is severed into discrete times and areas strongly reducing the extent to which these data inform stock depletion.

6. Recommendations

Conduct simple tests of model overparameterization.

Having developed a range of spatial multi-fleet operating models, it is my experience that models of the complexity presented here may not robustly estimate variables of management interest such as $F_{current}/F_{MSY}$ and $B_{current}/B_{MSY}$. This can be tested relatively easily by using the data-generation facility of Stock Synthesis to produce simulated fishery data in order to determine how well the model can retrieve known parameter values. It should be noted that such a test is likely to offer an optimistic evaluation of assessment performance since the same model structure is used to simulate the data. This is a 'first test' for overparameterization and will not help to diagnose problems due to differences between the real observation processes and dynamics and those assumed by the assessment model. Other tests include variability in the estimation of management reference points from different 'jittered' starting locations.

Include a more comprehensive evaluation of structural uncertainty: examine the marginal effect of adding different levels of complexity (seasonality, multiple fleets).

It would be instructive to see future assessments carried out in parallel with simple approaches that are quick and easy to apply. For example, delay-difference models or age-structured production models. For example a model assuming annual population and fishery dynamics, similar selectivities (1) among all longline fleets and (2) all other fleets and fitted to a single relative abundance index such as the Japanese longline fleet.

Identification of clear management objectives including target and limit reference points

The authors have done remarkably well at assessing a stock without established yardsticks with which to evaluate stock status and exploitation rate. It is paramount that managers establish target and limit reference points for the WCNPO striped marlin stock. Without a clear statement of what managers wish to achieve and avoid, it is not clear how to draw conclusions regarding the stock, make recommendations or develop management strategies. Currently, progress in the development of the quantitative tools with which to inform management may be hamstrung by a lack of clarity in objectives.

Greater clarity and rigor in the presentation of CPUE standardization methods

Given their importance for stock assessment, future CPUE standardization papers should endeavour to meet fundamental standards of scientific publication such as the clear description of a reproducible method and a transparent account of implicit assumptions (so that it is clear whether these are aligned with the assessment). The CPUE data may be confidential but at the very least the methods should be described in such a way that they may be applied to other CPUE datasets. Future standardization approaches should include maps of spatial coverage, graphs of the index itself (with an expression of uncertainty) and a suite of sensitivity analyses. Importantly the equation $I_t = \ldots$ should be included where I is the derived index and t the time subscript. While not preferable to annotated equations, a simple solution that could help meet this requirement would be for each standardization paper to include an Appendix of the computer code used to conduct the standardization.

A greater degree of theoretical consistency between CPUE standardization models and the assessment

Several CPUE GLM models assume time x area interactions or year x season interactions that imply population dynamics that are not accounted for by the spatially aggregated model. The model should avoid the use multiple regional abundance indices of equal prior weight (*e.g.* JP_DWLL1-3). If time x area interactions must be accounted for, this should be limited to a single standardization that models these interactions simultaneously and produces a single index weighted by the magnitude of regional abundance.

A greater degree of theoretical consistency among CPUE standardization models

We can see that year x season interactions, time x area interactions, sea surface temperature and set depth are all important factors in standardization of striped marlin CPUE. Some of these characteristics such as set depth are particular to specific gears. However where possible the methods should be consistent to avoid contradictory assumptions regarding the dynamics of the same stock.

Provision of CPUE data to increase the transparency of standardization and allow for cross-evaluation of methods

I acknowledge the difficulties in making publicly available, the fine-scale commercial catch and effort data with sufficient covariate data to investigate defensible standardization methods. For example, Japanese longline trip-level data at the 1x1 degree resolution with covariate hook-per-basket information, species composition, sea conditions, *etc*. However, since the derived relative abundance indices are critical to the assessment it is highly desirable that stock assessment scientists are provided with the basis to determine standardization procedures that may provide more consistent information about relative abundance across fleets.

In many stock assessment settings there are two central reference points, stock level (depletion, underfished / overfished) and exploitation level (underfishing / overfishing). I find the current *status quo* worrisome; that stakeholders arrive at the assessment with input data (the indices) that may largely predetermine one of these two key reference points (depletion). In my view, CPUE standardization is a critical part of stock assessment (interpreting fishery data in terms of abundance trends) and should be reviewed and investigated in the assessment meeting (perhaps this already occurs). It would, for example, be highly desirable for scientists involved in index calculation to attend the meeting with working code and allow the assessment group to interactively investigate the sensitivity of derived indices to different assumptions.

Simulation evaluation

It is important to understand where the trade-off lies in terms of assessment complexity versus reliability of management decision making. Consider the following hypothesis that may be tested by a multi-fleet operating model with some spatial population dynamics: quota recommendations derived from a simple one-stock, two fleet (longline and other) delay-difference model provides comparable management performance (assuming we have agreement on management objectives) to those of a more complex multiple-fleet, fully age-structured SS3 model (such as that applied here). This may well be rejected, but in other simulation settings simple assessment methods have provided a comparable and sometimes better basis for decision making whilst being more straightforward to apply and review. More complex assessments such as that applied here may offer detailed insights into the past characteristics fishing and the population but is that the objective of stock assessment? From my perspective it is important to know whether the additional complexity of such an assessment can be expected to offer practical benefits in terms of managing the stock.

Simulation evaluation also provides a basis for establishing management strategies that are robust to uncertainties of the population and fleet dynamics. For example, the regional abundance indices described may be valuable in defining spatial operating models to evaluate the magnitude of the single mixed WCNPO region and devise management procedures that perform well subject to a range of credible spatial hypotheses (seasonality in migration, incomplete stock mixing, *etc.*).

Spatial simulation models may provide a more objective basis for eliminating candidate abundance indices and determining the potential value of collecting other sources of data.

Consider a simpler base-case assessment model

It may be the case that a similar model with annual structure inferred by a two fleets (longliners and other fleets) fitted to a single relative abundance index (Japanese longline offshore) could provide credible management advice. The scientists involved in the design of the assessment are experts with a very good reputation for high- quality work. However there is simply not enough detail presented in the document to be sure whether the model is overly complex and capable of producing spurious predictions. My suspicion is that the assessment is simply asking too much from the data that are available.

Note that if the current assessment were to include many of the recommendations made here the document would be much larger and take substantially more resources to produce. This is also a problem associated with more complex assessment models; they pose difficulties for both the presentation and review process due to the manifold permutations of assumptions that may determine model behaviour in a complex way. If a comprehensive review is a key requirement of a stock assessment, a model of greater simplicity may be desirable from a practical standpoint.

Avoid using regional abundance indices to infer population-wide stock dynamics.

Instead, fit the base-case model to a single index of abundance derived from a fleet with good spatial coverage and important covariate information (in this case this may be the Japanese offshore longline fleet) and use other indices for sensitivity analysis.

Where possible, avoid the division of indices into temporal blocks to account for changes in fishing that may be incorporated in CPUE standardization.

Appendix 1. Bibliography

BILLWG. 2011b. Report of the Billfish Working Group Workshop. International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean. 24 May – 1 June 2011. Chinese Taipei, ROC.

Campbell, R.A. 2004. CPUE standardisation and the construction of indices of stock abundance in a spatially varying fishery using general linear models. Fish. Res. 70(2–3): 209–227.

Carruthers, T.R., McAllister, M.K. and Ahrens, R.N.M. 2010. Simulating spatial dynamics to evaluate methods of deriving abundance indices for tropical tunas. Can. J. Fish. Aqua. Sci. 47: 1409-1427.

Kadane, J.B., Lazar, N.A. 2004. Methods and criteria for model selection. J. Amer. Statistical. Assoc. 99: 279-290.

Kanaiwa, M., Kimoto, A., Yokawa, K. 2011. Standardized CPUE of striped marlin in north western Central Pacific Ocean by using GLM. ISC/11/BILLWG-2/06.

MAR&SWOWG. 2007. Report of the marline and swordfish working group joint workshop. International Scientific Committee for Tuna and Tuna-Like Species in the North Pacific Ocean. Annex 8. http://isc.ac.affrc.go.jp/pdf/ISC7pdf/Annex_8_ISCMARSWOWG_Mar_07.pdf

Miyake, P.M. 2004. A brief history of the tuna fisheries of the world. *In* Second Meeting of the Technical Advisory Committee of the FAO Project Management of Tuna Fishing Capacity: Conservation and Socio-economics, Madrid, Spain, 15–18 March 2004 *Edited by* W.H. Bayliff, J.I. De Leiva Moreno, and J. Majkowski. Food and Agriculture Organization of the United Nations, Rome, Italy. pp. 23–50.

Piner, K. Lee, H.H. 2011a. Meta-analysis of striped marlin natural mortality. ISC/11/BILLWG-1/10.

Piner, K. Lee, H.H. 2011b. Correction to Meta-analysis of striped marlin natural mortality. ISC/11/BILLWG-2/08

Piner, K.P., Conser, R., and DiNardo, G. 2006. Stock status of striped marlin in the north Pacific Ocean in 2005. Marlin and Swordfish Working Group Joint Meeting. 20-22 March, 2006, La Jolla, California, U.S.A. ISC/06/MARWG&SWOWG/07.

Piner, K.P., Conser, R., DiNardo, G., and Brodziak, J. 2007. SS2 sensitivity runs for striped marlin assessment WG 2007. Marlin and Swordfish Working Group Meeting. 19-26 March, 2007, Taipei, Taiwan. ISC/07/MARWG&SWOWG-1/02.

Sun, C.-L., Hsu, W.-S., Chang, Y.-J., Yeh, S.-Z., Chiang, W.-C., and Su, N.-J. 2011a. Reproductive biology of female striped marlin (Kajikia audax) in the waters off Taiwan (preliminary). Working paper submitted to the ISC Billfish Working Group Meeting, 19-27 January 2011, Honolulu, Hawaii, USA. ISC/11/BILLWG-1/11: 14p. Available at: http:// isc.ac.affrc.go.jp/ pdf/BILL/ ISC11_BILL_3/ ISC11BILLWG-3_WP01.pdf or http:// isc.ac.affrc.go.jp/ reports/ bill/ bill_2011_1.html1

Sun, C.-L., Hsu, W.-S., Chang, Y.-J., Yeh, S.-Z., Chiang, W.-C., and Su, N.-J. 2011b. Reproductive biology of male striped marlin, Kajikia audax, in the waters off Taiwan. Working paper submitted to the ISC Billfish Working Group Meeting, 24 May-1 June 2011, Taipei, Taiwan. ISC/11/BILLWG-2/09: 17p.

Sun, C.-L., Hsu, W.-S., Chang, Y.-J., Yeh, S.-Z., Chiang, W.-C., and Su, N.-J. 2011c. Age and growth of striped marlin (Kajikia audax) in waters off Taiwan. Working paper submitted to the ISC Billfish Working Group Meeting, 19-27 January 2011, Honolulu, Hawaii, USA. ISC/11/BILLWG-1/09: 15p. Available at: http://isc.ac.affrc.go.jp/pdf/BILL/ISC11_BILL_1/ISC11BILLWG1_WP09.pdf or http://isc.ac.affrc.go.jp/reports/bill/bill_2011_1.html1

Sun, C.-L., Su, N.-J., Yeh, S.-Z., Chang, Y.-J. 2011d. Standardized catch-rates for striped marlin (Kajikia audax) for Taiwanese distant-water longline fishery in the North Pacific Ocean for 1967-2009. ISC/11/BILLWG-1/07.

Yokawa, K. 2005. Standardized CPUE of striped marlin caught by Japanese (High Seas) large-mesh drift fishery in the north Pacific for the periods between 1977 and 1993. ISC\05\MAR&SWOWG\18.

Walsh, W.A., Lee, H.H. 2011. Standardization of Striped Marlin Kajikia audax CPUE with Generalized Linear Models fitted to Pelagic Longline Observer Data from the Hawaii-based Fishery in 1995–2009. ISC/11/BILLWG-1/08.

Appendix 2. CIE statement of work

Attachment A: Statement of Work for Dr. Tom Carruthers

External Independent Peer Review by the Center for Independent Experts

Stock Assessment of Striped Marlin

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 an 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: Striped marlin (*Tetrapturus audax*) is one of six species of billfishes commonly harvested multi-nationally from commercial and recreational fisheries in the western and central Pacific Ocean regions. Fishery management requires high quality science to effectively manage and conserve our living marine resources, and the scientific peer review of stock assessments by external CIE expertise is an important process in the determination of best scientific information available. The Terms of Reference (ToRs) of the peer review are attached in **Annex 2**.

Requirements for CIE Reviewers: Three CIE reviewers shall conduct an impartial and independent peer review in accordance with the SoW and ToRs herein. CIE reviewers shall have expertise, working knowledge and recent experience in the application of fish stock assessment, mathematical modeling, and statistical computing. Scientists who are employed by or have significant interactions with the Inter-American Tropical Tuna Commission (IATTC) and the Western and Central Pacific Fisheries Commission (WCPFC), and the Secretariat of the Pacific Community (SPC), should not be considered as reviewers. Scientists associated with the ISC also should be excluded as reviewers. Each CIE reviewer's duties shall not exceed a maximum of 10 days to complete all work tasks of the peer review described herein.

Location of Peer Review: Each CIE reviewer shall conduct an independent peer review as a desk review, therefore no travel is required.

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

Prior to the Peer Review: Upon completion of the CIE reviewer selection by the CIE Steering Committee, the CIE shall provide the CIE reviewer information (full name, title, affiliation, country, address, email) to the COR, who forwards this information to the NMFS Project Contact no later the date specified in the Schedule of Milestones and Deliverables. The CIE is responsible for providing the SoW and ToRs to the CIE reviewers. The NMFS Project Contact is responsible for providing the CIE reviewers with the background documents, reports, and other pertinent information. Any changes to the SoW or ToRs must be made through the COR prior to the commencement of the peer review.

Pre-review Background Documents: Two weeks before the peer review, the NMFS Project Contact will send (by electronic mail or make available at an FTP site) to the CIE reviewers the necessary background information and reports for the peer review. In the case where the documents need to be mailed, the NMFS Project Contact will consult with the CIE Lead Coordinator on where to send documents. CIE reviewers are responsible only for the pre-review documents that are delivered to the reviewer in accordance to the SoW scheduled deadlines specified herein. The CIE reviewers shall read all documents in preparation for the peer review.

Desk Review: Each CIE reviewer shall conduct the independent peer review in accordance with the SoW and ToRs, and shall not serve in any other role unless specified herein. **Modifications to the SoW and ToRs can 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 CIE Lead Coordinator.** The CIE Lead Coordinator can contact the Project Contact to confirm any peer review arrangements.

Contract Deliverables - Independent CIE Peer Review Reports: 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) Conduct an independent peer review in accordance with the ToRs (**Annex 2**).
3) No later than **18 November 2012**, 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 CIE Regional Coordinator, and to Dr. David Die, CIE Regional Coordinator, via email to ddie@rsmas.miami.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.

18 October 2012	CIE sends reviewer contact information to the COR, who then sends this to the NMFS Project Contact
25 October2012	NMFS Project Contact sends the CIE Reviewers the report and background documents
1-16 November 2012	Each reviewer conducts an independent peer review as a desk review

18 November 2012	CIE reviewers submit draft CIE independent peer review reports to the CIE Lead Coordinator and CIE Regional Coordinator
1 December 2012	CIE submits the CIE independent peer review reports to the COR
7 December 2012	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) each CIE report shall completed with the format and content in accordance with **Annex 1**,
(2) each 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: 301-713-2363 ext 136

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:

Gerald DiNardo, Stock Assessment Program Leader (NMFS Project Contact)
Pacific Islands Fisheries Science Center
2570 Dole Street, Honolulu, HI 96822-2396
Gerard.DiNardo@noaa.gov Phone: 808-983-5397

Kevin Piner
Pacific Islands Fisheries Science Center
2570 Dole Street, Honolulu, HI 96822-2396
Kevin.Piner@noaa.gov Phone: 858-546-7003

Jon Brodziak
Pacific Islands Fisheries Science Center
2570 Dole Street, Honolulu, HI 96822-2396
Jon.Brodziak@noaa.gov Phone: 808-983-2964

Hui-Hua Lee
University of Hawaii, Joint Institute for Marine and Atmospheric Research
2570 Dole St., Honolulu, HI 96822
Huihua.Lee@noaa.gov Phone: 808-983-5352

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.

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

Appendix 3. Typographical, grammatical errors *etc*.

Page 17, third paragraph, fourth sentence: "Japan longline fleets were targeting predominantly albacore for canning and occasionally caught striped marlin at the surface waters, whereas harpoon fisheries operating in coastal waters of Japan directly targeted striped marlin were."

Page 20, third paragraph, second sentence: "Although some fisheries have catch data time series extending back to at least 1952 and model were developed in parallel that included this early data…"

Page 24, first paragraph, second sentence. Incomplete reference.

Page 29, last paragraph, second sentence. "More uncertain dataset due to…"

Page 39, paragraph 4, first sentence. "The BILLWG identified important sebsutuvuty runs…"

Page 39, paragraph 5, second sentence. "Two indices (2 time periods) were used to replace night indices (3 area times 3 time periods)..."