Report on the 2013 Assessments of Hogfish in the South Atlantic and Gulf of Mexico

by R.I.C. Chris Francis
123 Overtoun Terrace
Wellington 6021
New Zealand

September 2014

Prepared for

The Center for Independent Experts

Executive Summary

This is a desk review of the 2013 stock assessment report for hogfish in the south Atlantic and Gulf of Mexico. The computer program Stock Synthesis was used to assess three stocks: West Florida (WFL), East Florida including the Florida Keys and Dry Tortugas (FLK/EFL), and Georgia through North Carolina (GA-NC).

The hogfish assessment team had a daunting task. The available data are weak, which made the assessment particularly sensitive to the many assumptions used in compiling the data and structuring the model. Moreover, a big proportion of the changes in stock abundance may have occurred in the years preceding the period with data. Also, Stock Synthesis is rather complex, with many options and structures, some of which are easily misunderstood because they are not well documented. Although the assessment team generally did a good job in setting up this first hogfish assessment in Stock Synthesis I found sufficiently many causes for concern that I could not be confident that the assessment results were robust, and thus that the assessments constitute the best scientific information available.

Issues of concern for all three stocks include the construction and weighting of length compositions; the choice of initial equilibrium catches and growth CVs (coefficients of variation); and the treatment of uncertainty in recreational catches. It is not yet determined how serious these issues are for the hogfish assessments, but collectively they certainly have the potential to be serious. For the FLK/ELF stock, I was concerned about the assumption, which has profound implications for the assessment, that growth was the same as in WFL. The basis and support for a default selectivity assumption also have important consequences.

In considering future research I recommend that

- Priority be given to
  - improving age and length sampling of hogfish fisheries (with an emphasis on ages), and
  - fishery-independent surveys for GA-NC
- Consideration also be given to
  - rationalizing the calculation of biomass indices, and
  - better identifying the location of the stock boundary between east Florida and North Carolina

For the next hogfish assessment I recommend

- That particular attention be paid to three data sets:
  - length compositions (their construction, selection, and weighting),
  - initial equilibrium catches, and
  - (the treatment of uncertainty in) recreational catches;
- That two hypotheses be reconsidered:
  - default selectivities, and
  - the choice of growth parameters by stock;
- The use of
  - a likelihood profile on R0,
  - a sensitivity analysis for steepness, and
  - the incorporation of uncertainty in the projection results.

1. Background

This report reviews, at the request of Northern Taiga Ventures, Inc. (see Appendix 2), the 2013 assessment for hogfish (*Lachnolaimus maximus*) in the south Atlantic and Gulf of Mexico using the Assessment Report and other supporting documents (Appendix 1). This review, part of the SEDAR 37 process (http://www.sefsc.noaa.gov/sedar/Index.jsp), is "responsible for ensuring that the best possible assessment is provided", and is intended to "provide guidance" to the Southeast Fisheries Science Center (see Appendix 2).

2. Reviewer's Role

The role of the present reviewer was to read the documents provided (listed in Appendix 1), as well as some associated documents (e.g., Stock Synthesis documentation and some papers referred to in the provided documents), and write an independent peer review report in accordance with his Terms of Reference (given in Annex 2 of Appendix 2).

3. Findings

I will first discuss four matters that seemed to me to be of particular importance in these assessments, and then present my findings in relationship to each of the eight Terms of References (TORs) for the review, as given in Annex 2 of Appendix 2. In some cases I found it difficult to decide to which part of the TORs I should attach a comment, because several parts seemed equally appropriate. Rather than repeat comments in several places I have usually arbitrarily assigned each comment to one place. In what follows, the initials "AR" will be used to identify references to parts (e.g., sections, tables, figures) of the Assessment Report.

In reading these findings it is important to understand that the Assessment Report treated hogfish in the south Atlantic and Gulf of Mexico as consisting of three stocks: West Florida (WFL), East Florida including the Florida Keys and Dry Tortugas (FLK/EFL), and Georgia through North Carolina (GA-NC). These stocks were assessed separately, using very similar model configurations, but different data, for each stock.

3.0 Four matters of importance

3.0.1 Initial depletion

For all three stocks, the SSB (spawning stock biomass) is estimated to have changed more before the assessment period (1986–2012) than during it (Table 1). Thus it is important to understand how the assessment models estimate initial depletion (SSB_{1986} as a percentage of SSB_{Virgin}), and how robust these estimates are. If these estimates are badly wrong then the whole assessment is undermined.

Table 1: Base model estimates of depletion at the beginning (1986) and end (2012) of the period covered by the assessment models (tabulated figures were calculated from AR Tables 11.2.4.1-3 and 11.2.7.1-3).

	Depletion (SSB as % of virgin SSB)	
Stock	1986	2012
WFL	34	50
FLK/EFL	9	8
GA-NC	43	21

The assessment models' calculation of initial depletion is rather complex, but five main groups of estimated parameters are involved. The first three – (log)R0, the InitF parameters (one for each fishery), and the selectivity parameters – are used to set up an equilibrium age structure for 1986, and this structure is then perturbed using the other parameters – SR_R1_offset and the Early_InitAge parameters (for ages 1-20). In these assessments these last two groups had little influence (because they were almost all close to zero – see AR Tables 11.1.4.1-3). Further, given R0 and the selectivities, the InitF parameters are pretty much determined by the user-provided initial equilibrium catches. Thus it seems that, roughly speaking, amongst the estimated parameters, it is R0 and the selectivity parameters that determine the estimated initial depletion. As to the data sets that will most influence estimates of these parameters, clearly the initial equilibrium catches will be important and so, I assume, will be the length compositions (these will obviously affect the selectivities, and for fisheries with asymptotic selectivity the smaller the proportion of large fish in early catches, the greater the initial depletion). A likelihood profile on R0 would help to identify the relative influences of each data set on this parameter, and thus on initial depletion.

3.0.2 Default selectivity assumptions

In all three assessments some fishery or survey selectivities were either forced to be domed (by setting parameter 6 of the double normal selectivity to -15), and some were forced to be asymptotic (by setting the same parameter to +15). In deciding what selectivity shape to use, the authors of the Assessment Report seem to have taken the view that the default assumption is that it be asymptotic. That is, we should assume a selectivity is asymptotic unless we have information to the contrary. This assumption is not explicitly stated, but I infer it from text like "no evidence exists to support a dome-shaped relationship" (AR Section 11.1.3.6) and the suggestion that the sensitivity analyses in which no selectivities were forced to be asymptotic are presented "for reference only" (AR Section 11.2.7.2.5) (I take this last comment to mean that these sensitivities should be excluded from those defining the range of uncertainty in the assessments).

I think it is important to be absolutely clear in the Assessment Report about whether the basis for this default selectivity assumption is meant to be scientific or precautionary. My view (explained in the next two paragraphs) is that it has little scientific support but could plausibly be adopted as precautionary.

From a scientific point of view, it is clear that a fishery which does not catch the largest fish in a population (either because of the characteristics of the gear or the spatio-temporal location of the fishery) must have a domed selectivity. However, it does not follow that a fishery that *does* catch these largest fish *must* have an asymptotic selectivity. All we can say is that this selectivity must be non-zero for these fish, and the selectivity *may* be asymptotic. In the information presented in the Assessment Report I could see no scientific justification for forcing any selectivity to be asymptotic (and I note the relevance of a recent paper by Waterhouse et (2014)). Further, this forcing was unsupported by the data. This can be seen in two ways: (a) an examination of the residuals for length compositions associated with forced-asymptotic selectivities shows that those for the largest length bins are mostly negative (e.g., consider plots for the commercial hook and line fisheries – AR Figures 11.2.1.3.4, 24, & 40); and (b) removal of the asymptotic restriction in some sensitivity analyses "In all cases ... led to estimation of a fully dome-shaped function" (AR Section 11.2.7.2.5).

It is often the case in stock assessments that allowing the model to determine whether or not selectivities are asymptotic will produce a less conservative (i.e., more optimistic) estimate of depletion than if at least one selectivity is forced-asymptotic. This was certainly true for the present assessments (Table 2). Some people argue that these less conservative assessments are unsound because some of the stock biomass is "cryptic" (i.e., theoretically present, but unselected by the fisheries and surveys). The view is taken that the appropriate, precautionary, approach is to avoid this so-called cryptic biomass by forcing at least one selectivity to be asymptotic. If this view is the basis of the authors' decision to force some selectivities to be asymptotic then this should be explicitly stated, both when the assumption is made (AR Section 11.1.3.6) and also when results are presented (to alert readers to the possibility that actual stock status may be more optimistic than is estimated).

Table 2: The effect on estimates of initial depletion of removing (in sensitivity analyses No_flattop) the base model restriction that some selectivities were forced to be asymptotic (base values from Table 1; No_flattop values from AR Figures 11.2.7.5.1-3).

	Initial depletion (SSB_{1986} as % of virgin SSB)	
Stock	Base	No_flattop
WFL	34	47
FLK/EFL	9	21
GA-NC	43	66

3.0.3 The role of the growth parameters

The same growth parameters (describing mean growth [i.e., mean length at age] and its variability) were used for all three stocks. To understand the implications of this decision we need to consider (a) the extent to which it is consistent with the available data, and (b) the roles played by the growth parameters in the stock assessment models.

The mean growth function was calculated using only WFL data. This function also seems reasonable for GA-NC, for which the limited data available are not obviously inconsistent with the WFL function, but it looks obviously wrong for FLK/EFL, where observed mean lengths at age are substantially less than those for WFL (AR Figure 5.5.4.1). The decision to use the WFL growth function for FLK/EFL was based on the assumption that the substantially lower lengths at age in FLK/EFL are simply a result of a higher exploitation rate, rather than any real ("genetic") differences in growth (AR Section 5.5.4). Here it seems that a distinction is being made between what one might call the "real" (or perhaps "intrinsic") growth function for FLK/EFL (that which would occur without fishing) and the "current" growth function (which is different because of fishing), and the argument is that we should use the "real" function (assumed equal to the WFL function) in the assessment model. As to variability in growth, it was assumed that the CV (coefficient of variation) of length at age was equal to 0.2 for all ages and stocks. Where this value came from is not explained. It appears only in the assessment Control Files (AR Section 12) and is clearly inconsistent with the data plotted for WFL and GA-NC in AR Figure 5.5.4.1 (for example, it implies that 95% of 15 year old fish would lie between about 42 cm and 98 cm, whereas the plotted data appear more consistent with a CV of 0.1 or less). Note that these CVs can, and should, be estimated by the software used to estimate the mean growth curve.

The growth parameters have two distinct roles in the assessment models. One obvious role is to convert from numbers at age (the models' accounting system) to biomass (the weight-length relationship is also used in this conversion). In this role they have an effect on the estimated productivity of the stock, as measured by various biological reference points (BRPs) such as $F_{30\%}$. I can appreciate that an argument may be made that BRPs should be based on "real", rather than "current" growth. However, compared to natural mortality and steepness, the influence of growth parameters on BRPs is typically small. A second, and much less obvious, role for growth parameters is in the fitting of length compositions. The likelihood of observed length compositions is calculated by comparing them to expected length compositions, which are obtained by converting the model's expected numbers at age using the growth parameters. For the FLK/EFL assessment this conversion will have been badly wrong, because the WFL growth function is inconsistent with the relationship between age and length at the time that the length composition data were collected. This inconsistency will also badly affect the likelihoods for the conditional age-at-length data. I note that the FLK/EFL assessment results changed substantially – initial depletion changed from 9% to 80% – in the sensitivity run that used mean growth based on data from this stock (see run AltGrowthFx in AR Figure 11.2.7.2.8.2). [As an aside, I note that the estimate of 80% initial depletion is inconsistent with the view that this stock has long been "severely overfished" (AR Section 3) but suggest that, rather than being a reason to avoid using stock-specific growth, it is evidence of the difficulty of obtaining a robust estimate of initial depletion from this model.] I'm guessing that when the assessment team decided to use WFL growth parameters in the FLK/EFL model they were focussing on the first of the above two roles for these parameters and overlooked the negative consequences associated with the second role.

The fact that the growth CV parameters were too high will adversely affect the fitting of both the length composition and conditional age-at-length data for WFL and GA-NC, though I am not sure by how much. It will make the expected length compositions (and the expected conditional distributions of age at length) markedly wider than they should be.

3.0.4 Data weighting

The weight given to each data point in a stock assessment model is determined by a measure of the assumed size of the error associated with that point: typically a CV for biomass indices, and a sample size for composition data. These weightings are important because they can substantially affect both model outputs and all statistical inference from the model (Francis 2011). Likelihood profiles illustrate how different data sets favour higher or lower estimates of the parameter being profiled (e.g., AR Figure 11.2.4.8 shows that the length data favour a higher estimate of steepness than did the age data; Francis 2011 gives another example of this phenomenon in his figure 1). If we change the data weighting, we change the balance between the different data sets, and thus change the parameter estimate. Statistical inferences from assessment models (e.g., estimation of standard errors [SEs], whether from the inverse Hessian, a bootstrap, or a likelihood profile) are valid only if the assumed CVs and sample sizes are consistent with the actual error in the data.

I was concerned by the lack of discussion of data weighting in the Assessment Report. My concern is primarily that the length composition data are probably over-weighted. The sample sizes used for these data appear to be actual sample sizes (with a maximum of 200). But *effective* sample sizes for length compositions (i.e., those consistent with the error in the data) are known to be smaller, often much smaller, than *actual* sample sizes because of what Pennington & Vølstad (1994) called intra-haul correlation (i.e., fish caught in the same haul are typically more alike in length than those caught in different hauls). Stock Synthesis provides a method for calculating effective sample sizes (using the 'effN' values in plots like AR Figure 11.2.1.3.1), though this seems not to have been used in the present assessments (e.g., I note that effN < N in all 7 composition data sets for FLK/EFL [see AR Figure 11.2.1.3.35]). However, this method ignores the correlations in composition data and consequently overestimates effective sample sizes (this is illustrated in table 4 of Francis (2011), in which sample sizes based on 'effN' values [method TA1.1] are 5-24 times higher than those calculated following Pennington & Vølstad (1994) [method TA1.8]).

3.1 TOR 1: The Data

Evaluate the data used in the assessment.

The information provided to me was insufficient for a thorough evaluation of the data used in the assessment.

With regard to the fishery data, I imagine the Assessment Report was written with a local audience in mind, and so no need was felt to describe the various data collection systems. I am now aware that there are two parallel systems for collecting recreational data (MRFSS and MRIP) but I have no idea as to how they differ in terms of the data they collect, and their collection methods, and so find it impossible to assess their absolute or relative reliability. Similar comments apply to two systems for collecting commercial fisheries data: by trip ticket and logbook. Having said this I should note that fishery data collection systems are usually, in my experience, rather complex, often containing many spatial, temporal, and data-type anomalies, so it is difficult for an outsider to fully understand their strengths and weaknesses.

All the abundance index data series were documented to some degree in the Background Documents, though the level of detail provided was very variable. Only the Reef Visual Census surveys were thoroughly documented (via Smith et al., 2011). One thing lacking from the documentation of the model-based estimates was an exploration of the factors/variables deemed to be significant (see, e.g., Bentley et al. 2012). If, for example, area is found to be significant, it is useful background information to know which areas have higher or lower catch rates, and how strong the effect is (do mean catch rates in different areas differ by 10% or 100%?). The plausibility of these effects is sometimes a useful diagnostic. I would have liked more information about the treatment of year x factor interactions in Background Document 37-12. Often such interactions are avoided because of the difficulty of devising an overall year effect. How that difficulty was avoided is of great interest, but was unclear to me.

a) Are data decisions made by the assessment panel sound and robust?

I agreed with many of the data decisions made by the assessment panel. However, there are some decisions that I think were unsound. Moreover, I do not agree with the common point of view, apparently shared by the panel, that (almost) all available data should be presented to the model, regardless of how sound they are. I realise that it can be politically difficult to omit inadequate data, because this may offend those who collected the data, and the institutions that funded them. However, it is my fear that unrepresentative data can easily mislead stock assessment models, whose ability to distinguish signal from noise is not great, particularly in assessments like these where the signals are so weak. (When it is impossible to resist the pressure to include inadequate/suspect data I have found that a good strategy is to include them only in sensitivity analyses, or model runs that provide an alternative to the base assessment.)

For reasons given in Section 3.0.3 I think it was a serious mistake to use the WFL mean growth parameters in the FLK/EFL assessment, and also to use growth CVs that were inconsistent with the available data for at least two of the stocks (WFL and GA-NC).

A much more minor growth-related matter that caught my attention was the decision to include only data from the life history studies in the estimation of growth parameters, apparently because the otolith readings for all other available data were not made using the validated methodology of McBride & Richardson (2007) [Supplementary Document S37_RD04]. This decision would be standard, and sound, in an age and growth study, but it seemed odd in the context of this stock assessment because the excluded data apparently included all the conditional ages at length that were used in the assessment. Surely if the ages in these data were not sound enough to estimate a growth curve for use in the assessment then they should not have been included as observations in the assessment model. A simple qualitative check of the quality of these data is to plot residuals (observed length minus the growth curve mean length at age) against age using different plotting symbols for the two data sets (the life history data used to estimate the growth curve, and the assessment observations).

I was concerned about all of the length compositions used in the assessments, except for those from the Reef Visual Census (RVC) surveys. It was only for these surveys that I could find any description of how the length compositions were constructed. Without this information I am left with the strong suspicion that these compositions were simply raw data. This is inappropriate for fishery compositions because fishery catches are typically very heterogeneous. In particular, the length compositions of individual catches often vary in systematic ways, depending on factors such as the area fished, time of year, or vessel type. Thus a common way to obtain a length composition that might be representative of the catch for a given year is to (a) find out which factors most affect catch length compositions, (b) make sure to sample across these factors (i.e., stratify the catch), and (c) use these factors to scale up the length samples (e.g., if samples are collected for each quarter of the year, they should be weighted by the proportion of the total catch that comes from each quarter). Such a scheme highlights the necessity of discarding as unrepresentative data from years in which there was not adequate sampling in all strata. The sample sizes for many of the individual length compositions in these assessments were clearly too small to be considered representative. With stratified surveys it is conventional to use stratum area in weighting the compositions, but I saw no reference to such weighting in the documents provided (except for the RVC surveys).

Too little attention appears to have been paid to the initial equilibrium catches provided for each fishery. These are important because they affect the calculation of the initial (1986) age structure, and thus the estimation of initial depletion (see Section 3.0.1). In essence, the model assumes that the effect on the stock of all pre-1986 fishing is the same as if these initial equilibrium catches had been taken every year before 1986 for as long as it would take for the population to reach an equilibrium. Thus I was surprised to find that, with no discussion in the text, all but one of the initial equilibrium catches were set equal to the 1986 catch, rather than being based on the pre-1986 catches (the exception was Rec_HL in the GA-NC assessment). The 1986 catches were sometimes markedly less than most preceding ones (e.g., both recreational fisheries for FLK/EFL), and sometimes markedly more (e.g., Rec_Spear for WFL).

I am doubtful about the inclusion of three of the abundance indices – Video and SEAMAP for WFL, and RVCKeys for FLK/EFL – though I acknowledge that excluding them may not have much effect on the assessments. For the Video surveys my concerns are (a) the many changes over time in the equipment deployed and (b) doubt about whether the quantity measured (maximum number of hogfish in a single video frame) is proportional to abundance. For SEAMAP I was concerned that of the 5 zones used in the analysis, all were occupied in only three of the five years. For RVCKeys, I was unconvinced about the assumption that there was a step change in catchability in 2000. This is explained in AR Section 11.1.4 as being "to model a change in catchability reflecting updates to the RVC methodology and increases in precision". In Smith et al. (2011) I noted that the survey design/methodology was altered several times in different years, but I found no suggestion that the changes in 2000 could have more than doubled the survey catchability. The postulated change in catchability in 2000 might have been more convincing if evidence were presented of other species whose abundance index had jumped up in 2000. As to the reference to "increases in precision", (a) the undoubted improvement in precision over time seems to me to have been caused by a gradual increase in sampling intensity (I found that a plot of log(CV) vs *n* from table 1 of SEDAR37-09 is fairly linear), and (b) why should there be a connection between changes in catchability and precision?

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

The Assessment Report acknowledged major data uncertainties associated with recreational catches, stock structure, and, for FLK/EFL, growth rates and conflicting abundance index trends. These sorts of uncertainties are common in stock assessments, and their levels in the present assessments are not out of the ordinary.

One source of uncertainty that did not seem to be explored or acknowledged concerns the distinction between hogfish and non-hogfish habitat, which was used in the construction of the commercial and recreational CPUE (catch per unit effort) indices. Records from non-hogfish habitat were excluded from the index calculations. It is sensible and reasonable to make this distinction, but it seems to me that the boundary between the two habitats is essentially arbitrary. What remains uncertain is the degree to which the CPUE indices might change if the arbitrary definition of the boundary were changed. It could well be that the CPUE indices are relatively insensitive to this boundary definition, but we won't know that until the sensitivity has been explored.

c) Are data applied properly within the assessment model?

The data were generally well applied within the assessment model, with two exceptions.

First, I was concerned about the weighting of the length composition data: this was not discussed in the Assessment Report, and I am concerned that these data appear to have been over-weighted, which could have substantial effects on the assessment results (see Section 3.0.4).

My second concern relates to the treatment of the undoubtedly high uncertainty about the recreational catches. I think it is almost always a mistake to assign high SEs to any catches in assessment models. To do so is to say to the model "we don't really know what the catches from this fishery were, please estimate them for us". This is a mistake because I can't see what information the model has to make these estimates. I was particularly concerned to see that the model's estimates of recreational hook and line catches for the GA-NC assessment were substantially higher than the "observed" catches (AR Figure 11.2.1.1.3). Thus this model assesses the stock using catches that were much higher than the data indicate. On what basis? I suggest that a better way to deal with uncertainty in catch histories is to use best-estimate catches in the base model and then run sensitivity analyses with alternative high and low catch histories. This approach provides an explicit measure (lacking in the current assessments) of the effect of catch uncertainty on all assessment outputs. Note that by assigning low SEs to the catches we are not intending to assert that these catches are well known; we are simply trying to find out what the stock status would likely be *if* these catches were correct. [Technical note. The "hybrid F" option used in these assessments is recommended in Stock Synthesis, as stated in the Assessment report (see AR Section 11.1.4), but *not* for the case when catch SEs are high (see section 9.3.14 in Methot 2012).]

The best way to construct alternative high and low recreational catch histories depends on the nature of the errors in the catches. If this is purely sampling error (arising because a small number of trips has been sampled) then we can assume the errors are independent between years and calculate a variance, and thus 95% confidence interval, for the total catch (for a given stock and fishing method) by summing the catch variances across years. Then the low and high catch histories could be constructed by scaling the base catch history so that its total spanned the 95% confidence interval. If the catch errors involve bias (e.g., biased low because some sectors of the fishery are never sampled) then we must try to put bounds on the likely size of that bias, and use these (as well as the sampling error) in constructing low and high catch histories.

The two extreme spikes in the recreational hook and line catch history for GA-NC (AR Figure 11.2.1.1.3) are also of concern. I am surprised that these were included despite the authors' acknowledgments (in paragraph three of AR Section 11.2.5) that (a) "such drastic year-to-year changes in landings" are "unlikely", and (b) the first spike caused the model to estimate an implausible biomass trajectory (AR Section 11.2.5). An important point about the effect of catches in assessment models is that, in most assessments, all that matters is that the total historical catch from each fishery, and any broad trends in that catch, are approximately correct. An error in the catch for a specific year is usually unimportant, as long as the total catch from all nearby years is about right. This is a good reason, when catches are highly uncertain, to present catch histories that are reasonably smooth.

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

For reasons given above (see beginning of Section 3.1) it is difficult for me to fully evaluate the reliability of the input data series in these assessments. However, my general impression is that, with one exception, these data series are, on balance, no less reliable than those used in many acceptable stock assessments for small stocks in other parts of the world. The exception is the length composition data sets with very low sample sizes. It's not possible to specify a minimum acceptable sample size because the sample structure is important too (e.g., a sample of 200 fish would not be acceptable if they all came from one trip, which contributed only a small percentage of the catch for that fishery). However, it is hard to believe that the many compositions with sample sizes less than 20 could be representative of the catch for a year.

3.2 TOR 2: Assessment methods

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

Evaluation of the methods used in the base assessments was greatly facilitated by the use of Stock Synthesis. I commend the authors for providing (a) the Stock Synthesis input files (AR Section 12), (b) tables of parameter estimates (AR Tables 11.1.4.1-3), and (c) many standard plots of outputs generated by the r4ss software. These allowed me to understand, in considerable detail, aspects of the data and model assumptions that were not clear from the text.

a) Are methods scientifically sound and robust?

The stock assessment program Stock Synthesis used in these assessments is widely used, well tested, and has a strong reputation, as evidenced by the recent special issue of the journal *Fisheries Research* (volume 142, 2013) that was dedicated to it. To use such a program is wise because it protects against coding errors and ensures that state of the art approaches to stock assessment are available. However, it does not imply that an assessment will be scientifically sound and robust, because this depends on the quality of the available data and many decisions made by the assessment team. Some decisions that have, or may have, undermined the present assessments are discussed under TOR 1 and the rest of TOR 2.

b) Are assessment models configured properly and used consistent with standard practices?

The assessment models are generally well configured, though it's difficult to comment in relation to "standard practices" because, in my experience, what is "standard" in stock assessments varies widely amongst institutions. There was one aspect of model configuration that I thought had a particularly strong bearing on these assessments.

The decision to force some selectivities in each assessment to be asymptotic was of key importance. In my view, this decision was "proper" if and only if (a) its basis was precautionary, rather than scientific, and (b) it is appropriate within the SEDAR system for a stock assessment team to make precautionary decisions (see Section 3.0.2). I make this latter point because in New Zealand (where I have worked) the invocation of precautionary arguments is explicitly outside the domain of assessment scientists.

c) Are the methods appropriate for the available data?

In general the methods used were appropriate for the available data. The only comment I would make is that more parameters were estimated than could be justified by the available data. In particular, I am sure that there was no point in estimating SR_R1_offset, and am dubious about most of the Early_InitAge parameters (as noted above, in Section 3.0.1, almost all these parameters differed only insignificantly from zero). Although estimating these parameters violates the principle of Occam's Razor, it does little other harm. However, I think it was definitely a mistake to try to estimate steepness (parameter SR_BH_steep), which is notoriously difficult, even in assessments with much better quality data than the present ones (Lee et al. 2012). In Section 3.5 I suggest a better way to deal with steepness.

3.3 TOR 3: Assessment findings

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

For all stocks I am not confident of the reliability of estimates of abundance, exploitation, and biomass, or of estimates of the status with regard to overfished and overfishing thresholds.

Much of my lack of confidence is associated with the length compositions. Abundance indices are not very informative in stock assessments unless they show a strong trend, which was not the case for any of the indices used in these assessments. Therefore much of the inference in these assessments must depend on the length compositions (I discount the third type of observation – conditional ages at length – because of small sample sizes for almost all years). However, I am dubious about both the construction and weighting of the compositions (see Section 3.1a,c). Further, inferences based on length compositions require sound growth parameters (see Section 3.0.3), which I think were arguably lacking for all three stocks, but particularly for FLK/EFL (see Section 3.1a). I note also the link between the compositions and initial depletion, which is important in all three stocks (see Section 3.0.1). Another concern is the initial equilibrium catches, some of which were clearly inappropriate (see Section 3.1.a). I don't know how much effect a different treatment of the composition data and more appropriate initial equilibrium catches would have on the estimates from these assessments, but I think it quite possible that they could cause a substantial change.

The GA-NC assessment is particularly problematic because there was only one abundance index and this was not well fitted by the model (AR Figure 11.2.1.2.16). I suggest that either the abundance index is reliable, in which case the stock assessment (with which it is inconsistent) is not; or it is not reliable, in which case an assessment of this stock is not possible. I note however, that better weighting of the composition data (see Section 3.1c) might allow a good fit to the abundance index.

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 stock recruit relationship was not well determined for any of the three stocks because the assessment period covered a relatively limited range of spawning biomass (AR Figures 11.2.4.1, 5, 9). This markedly limits the precision of estimates of steepness, and thus productivity and future stock conditions. However, I note that it is possible, via sensitivity analyses, to explore how uncertainty in the steepness of the stock-recruit relationship affects these estimates (see Section 3.5).

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?

I presume that "status determination criteria" refers to the F- and biomass-ratios in AR Tables 11.2.7.1-3 associated with the reference points of AR Section 11.2.8. I am not confident of the reliability of estimates of these ratios (for reasons given at the beginning of Section 3.3), and I am not aware of other indicators that are likely to be more useful.

3.4 TOR 4: Stock projections, rebuilding timeframes, and generation times

Evaluate the stock projections, rebuilding timeframes, and generation times, addressing the following:

For these assessments, projections were carried out and rebuilding timeframes were discussed, but generation times were not mentioned.

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

In the realm of stock projections, what constitutes "accepted practice" varies greatly from place to place. The methods used here were those provided by Stock Synthesis, which I am sure reflect accepted practice on the west coast of the U.S.A. (though not in New Zealand), and are certainly not inconsistent with the available data. I don't know whether they are considered to be accepted practice within SEDAR.

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

I am not aware of any aspect of the assessment model or its outputs that would make the Stock Synthesis projection methods inappropriate.

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

I found the projection results hard to interpret. This is no fault of the assessment team (although it didn't help that the columns in some of the tabular output seem to have been mislabelled [e.g., I think the columns labelled F40 in AR Tables 11.2.9.1-2 should have been FCurr] – I assumed that the labelling in the associated plots [AR Figures 11.2.9.1-3] was correct). It is simply a consequence of the rather complex multiple-pass projection methodology (see appendix B of Methot 2012), which differs substantially from what I am accustomed to and produces some counter-intuitive results (e.g., see the first paragraph of AR Section 11.2.9, and note that for GA-NC, projected SSBs were higher with FCurr than with FMSY [AR Figure 11.2.9.3] despite FCurr being higher than FMSY [according to AR Table 11.2.7.1.3]). It may be that the projections also confused the assessment team because the text discussing them seemed sometimes to be inconsistent with the associated plots (e.g., I think "within 5 yrs" in the last paragraph of AR Section 11.2.9 should read something like "in 6-8 yrs"; also, in the preceding paragraph, "9 yr" in last sentence seemed inconsistent with "15-20 yrs" in the third sentence). However, if the Stock Synthesis projection methodology is accepted practice within SEDAR then I must assume that those who need to be able to interpret outputs such as those presented here have developed sufficient familiarity with them to find them informative.

As to robustness and utility in inferring probable future conditions, no projection results can be more robust than the associated assessment, and, as explained in Section 3.3, I am not convinced that the assessment results are robust.

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

The projection results (as presented in AR Section 11.2.9 and associated tables and figures) included no presentation or discussion of uncertainty. For me, one of the weaknesses of the Stock Synthesis projection outputs (at least those presented in the Assessment Report) is that, although the calculations include some uncertainty (in the form of stochastic recruitment and catch implementation error), this is not reflected in the outputs. However, I would acknowledge that for many assessments (and certainly for the present ones) the uncertainty included in the projection calculations for a base model run is small compared to that between the base and plausible alternative models.

3.5 TOR 5: Uncertainties

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

Five methods of addressing (various types of) uncertainty were considered in the Assessment Report: bootstrap simulation results (AR Figures 11.2.7.1.1-15); asymptotic (i.e., inverse Hessian) estimates of parameter SEs (AR Tables 11.1.4.1-3); likelihood profiles on steepness (AR Figures 11.2.4.4, 8, 12); sensitivity analyses (AR Figures 11.2.7.2.1.1-8.3); and retrospective analyses (AR Figures 11.2.7.3.1-3).

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

For me, the least informative of the five methods were the likelihood profiles on steepness and retrospective analyses. The former were uninformative because (a) the data contain very little information about steepness (see Section 3.3d), and (b) the profiles are meaningless unless the data weightings are correct, which I don't think was true (see Section 3.0.4). For reasons given in Section 3.0.1 I think a much better choice for a parameter to profile would have been R0, with the aim being to understand which data sets were influencing the estimation of this parameter, and thus how robust the estimates of both R0 and initial depletion are. As to retrospective analyses, the problem with these is that when retrospective trends are found (as they were for WFL and GA-NC), it is very difficult to determine whether they are of concern (i.e., indicating some sort of model miss-specification) or simply arising from random patterns in observation error. Further, very similar trends can be caused by completely different types of model miss-specification (Mohn 1999).

I am dubious about asymptotic errors because I don't think the stock assessment situation is anything like asymptotic (what is meant by large sample size in an assessment with many sets of observations, some from complex sampling schemes?), and was surprised by the conclusion that "the asymptotic errors ... tended to produce similar estimates to the bootstrap parameter estimates" (AR Section 11.2.7.1) because my very limited comparison showed that these often differed by a factor of more than 2 (and not always in the same direction).

I think that a well-chosen set of sensitivity analyses usually provides the best description of uncertainty for stock assessments. A bootstrap approach is theoretically appealing but (a) as with a likelihood profile, the present results are probably misleading because composition data weightings (i.e., assumed observation errors) are wrong, and (b) my experience is that alternative model assumptions often cause a bigger change in model outputs than could be attributed to observation error (consider, e.g., the effect in the present sensitivity analyses of dropping the assumption that some selectivities must be asymptotic).

I offer some brief comments on the choice of sensitivity analyses and their interpretation:
– just removing 1 of the 8 biomass indices for WFL is unlikely to make much difference (AR Section 11.1.7.2);
– it would have been more informative to use alternative recreational catch histories (as suggested above, see Section 3.1c) rather than fiddling with SEs (AR Section 11.1.7.3);
– it would have been more informative to treat steepness in the same way as natural mortality – i.e., fix it in the base model and use bracketing values in the sensitivity analyses (the fixed and bracketing values could be taken as the 50th, 5th, and 95th percentiles of the prior) (AR Section 11.1.7.4);
– dropping the forcing of asymptotic selectivity (AR Section 11.1.7.5) is a key sensitivity;
– simply changing the conditional length at age data to unconditional makes sense only in the unlikely case that the fish can be considered a simple random sample from the catch; and are otolith sample sizes really big enough to estimate growth in FLK/EFL and GA-NC? (AR Section 11.1.7.7);
– the sensitivity analysis using stock-specific growth for FLK/EFL is a key result; for reasons given in Section 3.0.3 I don't agree with the suggestion that it "should not be viewed as a reasonable alternative, but useful for illustrative purposes" (AR Section 11.2.7.2.8).

A useful way to incorporate the uncertainty that was lacking from the projection outputs (see Section 3.4d) would have been to repeat the projections for a selected few of the sensitivity analyses.

b) Ensure that the implications of uncertainty in technical conclusions are clearly stated.

I am not sure that readers will appreciate the true range of uncertainty in the results from these assessments because (a) discussions of uncertainty seemed to focus mostly on the bootstraps results, with comparatively little attention paid to key sensitivity analyses, (b) there was no quantitative summary of the likely overall uncertainty (from bootstraps and sensitivity analyses) in key outputs and (c) no uncertainty was presented in the projection results.

3.6 TOR 6: Research recommendations

Consider the research recommendations provided 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 Assessment Report presented four research recommendations, which can be summarised as follows.

1. More growth, maturity, and fecundity information for FLK/EFL and GA-NC
2. Fishery-independent surveys for GA-NC
3. Improve age and length sampling of hogfish fisheries
4. Contribution of males to spawning reproductive potential

Of these, I think 3 and 2 are most likely to improve future assessments. With regard to 3, I would strongly emphasize the age data (age compositions are *much* more informative in a stock assessment model than length compositions) and comment that doubling the number of trips sampled has a much greater effect on precision than doubling the number of fish sampled per trip. There is a need to check that the ageing of the sampled fish is consistent with the validated method (see Section 3.1a) and to develop reliable methods to construct length compositions from these data (see Section 3.1a). With regard to 2, I note that GA-NC was the weakest of the three assessments because there was only one biomass index, and this was not well fitted by the model (see AR Figure 11.2.1.2.16).

I think that 4 is important, but perhaps more to inform fishery management (e.g., are special measures needed to protect males?) than to improve the stock assessment.

A comparative study of alternative methods of biomass index calculation would be useful. I was startled by the array of different methods used for indices in these assessments (markedly different methods are described in Background Documents 37-02, 37-05, 37-09, and 37-12) and wondered whether it was either necessary or desirable to use so many different methods. Perhaps some rationalization of methods would be possible. The study should include an investigation of sensitivity to habitat classification (see Section 3.1b) and the desirability of including year interactions (see beginning of Section 3.1) and, for the survey data, design-based methods (because they require fewer assumptions I think these methods are preferable unless demonstrated to be markedly less precise than the model-based methods).

Background Document 37-01 showed that fish from east Florida waters were genetically distinct from those from North Carolina waters, but was unable to be specific about the location of the boundary between these stocks. Samples from Georgia and South Carolina would be useful to check the stock assessment assumption that this boundary lies at the Florida-Georgia border.

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

I have no recommendations to make but would like to comment that I think the present review would have been much better informed, and thus more useful, had it involved participation in an assessment review meeting, during which both discussions with the assessment team and some additional model runs could have resolved some important uncertainties for me.

3.7 TOR 7: Improvements for next assessment

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

Three sets of data merit particular attention in the next assessment. For length compositions, improve their construction (Section 3.1a), and selections (i.e., dropping years with small [unrepresentative] sample sizes – see Section 3.1d), and improve their weighting (Section 3.1c). Initial equilibrium catches need to be better related to pre-1986 catches (Section 3.1.a). For recreational catches, use small SEs, smooth spikes, and construct alternative low and high catch histories (Section 3.1c) to use in sensitivity analyses.

Two other model assumptions are worth reconsidering: the default selectivity (Section 3.0.2), and the growth parameters (including CVs) used for each stock (Section 3.0.3).

Some other suggestions that might improve the next assessment are as follows.

- Use a likelihood profile on R0 to help understand the data sets affecting the estimation of this parameter, and thus the robustness of the assessments' estimates of initial depletion (Section 3.0.1).
- Use sensitivity analyses to investigate the effect of uncertainty in steepness, just as was done for natural mortality (Section 3.5a).
- Characterize uncertainty in the projections by repeating them for selected sensitivity analyses (Section 3.5a).

3.8 TOR 8: Review report

Prepare a Peer Review Report summarizing the Reviewer's evaluation of the stock assessment and addressing each Term of Reference.

You are reading the peer review report.

4. Conclusions and recommendations

The hogfish assessment team had a daunting task. The available data are weak: the abundance indices failed to show any strong trend; age data (often the second most informative data set) were very slight; and even the length data (typically a poor third in order of information content) were quite patchy. Moreover, a big proportion of the changes in stock abundance may have occurred in the years preceding the period with data (see Section 3.0.1). When the data are weak, assessments are more sensitive to the many assumptions that are made in compiling the data and structuring the model. The possibility of making an unfortunate assumption was particularly high in these assessments because they were the first for hogfish using Stock Synthesis, so the assessment team often could not rely on assumptions made, and ratified, in previous assessments. Furthermore, Stock Synthesis is rather complex, with many options, some of which are easily misunderstood (e.g., the assessment team's choice of initial equilibrium catches (see Section 3.1a) suggests that they may not have fully appreciated the role these played; this role is not well described in the Stock Synthesis User Manual).

Another complexity easily overlooked is the important role played by growth parameters in the fitting of length composition data (see Section 3.0.3).

Although the assessment team generally did a good job in setting up this first hogfish assessment in Stock Synthesis, I found sufficiently many causes for concern that I could not be confident that the assessment results were robust. This is particularly true for the FLK/ELF stock because of the profound implications of assuming the same growth as in WFL (see Section 3.0.3). For all stocks, I found reason for concern in issues such as the construction and weighting of length compositions (see Section 3.1a,c), the choice of initial equilibrium catches (see Section 3.1a) and growth CVs (Section 3.0.3), and the treatment of uncertainty in recreational catches (Section 3.1c). It could be that some or all of these issues turn out not to be important for the present assessments (e.g., I have found that changing the weighting of composition data makes a big difference to some assessments, but very little to others). It is their *potential* to cause substantial changes that makes me lack confidence in the robustness of these assessments.

Another issue that has the potential to undermine these assessments is the assessment team's default selectivity assumption (Section 3.0.2). Should this assumption not be deemed "proper" (see discussion in Section 3.2b) then all three assessments will be too pessimistic.

4.1 Recommendations

Amongst the Assessment Report's four suggestions for future research I recommend giving priority to the two concerning improve age and length sampling of hogfish fisheries (the focus should be on ages) and fishery-independent surveys for GA-NC. In addition, I recommend (a) research to rationalize the calculation of biomass indices, and (b) genetic sampling to better identify the location of the stock boundary between east Florida and North Carolina. (See Section 3.6a for more details).

For the next hogfish assessment I recommend that particular attention be paid to three data sets: length compositions (construction, selection, and weighting), initial equilibrium catches; and the treatment of uncertainty in recreational catches. Two hypotheses that should be reconsidered are those concerning default selectivities and growth. Other recommendations include a likelihood profile on R0, a sensitivity analysis for steepness, and a way of characterizing uncertainty in the projection results. (See Section 3.7 for more details).

5. References

Bentley, N.; Kendrick, T.H.; Starr, P.J.; Breen, P.A. 2012. Influence plots and metrics: tools for better understanding fisheries catch-per-unit-effort standardizations. *ICES J. Mar. Sci.* 69 (1): 84-88.

Francis, R.I.C.C. 2011. Data weighting in statistical fisheries stock assessment models. *Can. J. Fish. Aquat. Sci*. 68: 1124-1138.

Lee, H-H.; Maunder, M.N.; Piner, K.R.; Methot, R.D. 2012. Can steepness of the stock–recruitment relationship be estimated in fishery stock assessment models? *Fish. Res.* 125– 126: 254– 261.

Methot, R. D. 2012. User Manual for Stock Synthesis Model Version 3.24f. Updated October 3, 2012. NOAA Fisheries, Seattle, Washington.

Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. – ICES J. Mar. Sci. 56: 473–488.

Pennington, M.; Vølstad, J.H. 1994. Assessing the effect of intrahaul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, 50(3): 725–732.

Smith, S.G., Ault, J.S., Bohnsack, J.A., Harper, D.E., Luo, J., McClellan, D.B. 2011. Multispecies survey design for assessing reef-fish stocks, spatially-explicit management performance, and ecosystem condition. *Fish. Res.* 109: 25-41.

Waterhouse, L.; Sampson, D.B.; Maunder, M.; Semmens, B.X. 2014. Using areas-as-fleets selectivity to model spatial fishing: Asymptotic curves are unlikely under equilibrium conditions. *Fish. Res.* 158: 15-25.

Appendix 1 Materials Provided

The materials provided for the review fell into three categories: the assessment report; background documents; and supplementary documents.

Assessment Report

Cooper, W.; Collins, A.; O'Hop, J.; Addis, D. 2014. The 2013 Stock Assessment Report for Hogfish in the South Atlantic and Gulf of Mexico. Florida Fish and Wildlife Commission. 500 p.

Background Documents

SEDAR37-01 Seyoum S, Collins AB, Puchulutegue C, McBride RS, Tringali MD. 2014. Genetic population structure of Hogfish (Labridae: *Lachnolaimus maximus*) in the southeastern United States.

SEDAR37-02 Cooper W. 2014. Commercial catch per unit effort of Hogfish (*Lachnolaimus maximus*) from Florida Trip Ticket landings, 1994-2012.

SEDAR37-03 Cooper W. 2014. Recreational catch per unit effort of Hogfish (*Lachnolaimus maximus*) in the Southeast US using MRFSS-MRIP intercept data, 1991-2012.

SEDAR37-04 Cooper W. 2014. Relative index of abundance from visual order-of-magnitude REEF surveys applied to Hogfish (*Lachnolaimus maximus*) in the Southeast US, 1994-2012.

SEDAR37-05 Switzer TS, Keenan SF, McMichael RH Jr, DeVries DA, Gardner CL, Raley P. 2013. Fisheries-independent data for Hogfish (*Lachnolaimus maximus*) from reeffish video surveys on the West Florida Shelf, 2005-2012.

SEDAR37-06 Switzer TS, Fischer KM, McMichael RH Jr. 2013. Fisheries-independent data for juvenile Hogfish (*Lachnolaimus maximus*) from the annual FWRI SEAMAP trawl survey, 2008-2012.

SEDAR37-07 Switzer TS, Fischer KM, McMichael RH Jr. 2013. Fisheries-independent data for juvenile Hogfish (*Lachnolaimus maximus*) from the annual baitfish survey, 2002-2012.

SEDAR37-08 Switzer TS, Keenan SF, McMichael RH Jr, Fischer KM. 2013. Fisheries independent data for juvenile Hogfish (*Lachnolaimus maximus*) from polyhaline seagrasses of the Florida Big Bend, 2008-2012.

SEDAR37-09 Smith SG, Ault JS, Bohnsack JA, Blondeau J, Acosta A, Renchen J, Feeley MJ, Ziegler TA. 2013. Fisheries-independent data for Hogfish (*Lachnolaimus maximus*) from reef-fish visual surveys in the Florida Keys and Dry Tortugas, 1994-2012.

SEDAR37-10 Bachelor and Reichert. 2014. Summary information for Hogfish *Lachnolaimus maximus* seen on videos collected by the SouthEast Reef Fish Survey in 2010 – 2012 between North Carolina and Florida.

SEDAR37-11 Hiltz et al. 2014. Standardization of commercial catch per unit effort of Hogfish (Lachnolaimus maximus) from South Carolina Trip Ticket landings, 2012.

SEDAR37-12 McCarthy. 2014. Analysis of Hogfish data from Coastal Fisheries Logbook Program (CFLP).

SEDAR37-13 Collier. 2014. Standardization of commercial catch per unit effort of Hogfish (*Lachnolaimus maximus*) from North Carolina Trip Ticket landings.

Supplementary Documents

SEDAR37-RD-01 Collins, A.; McBride, R. 2008. Integrating life history, mating system, fishing effects, and habitat of hogfish, *Lachnolaimus maximus*, a harem spawning fish in the southeast U.S.

SEDAR37-RD-02 Collins, A.B.; McBride, R.S. 2011. Demographics by depth: spatially explicit life-history dynamics of a protogynous reef fish. Fish. Bull. 109:232–242

SEDAR37-RD-03	McBride, R.S.; Johnson, M.R. 2007. Sexual development and reproductive seasonality of hogfish (Labridae: *Lachnolaimus maximus*), an hermaphroditic reef fish. Journal of Fish Biology 71, 1270–1292
SEDAR37-RD-04	McBride, R.S.; Richardson, A.K. 2007. Evidence of size-selective fishing mortality from an age an d growth study of hogfish (Labridae : *Lachnolaimus maximus*), a hermaphroditic reef fish. Bulletin of Marine Science, 80(2): 401–417.
SEDAR37-RD-05	McBride, R.S.; Thurman, P.E.; Bullock, L.H. 2008. Regional variations of hogfish (*Lachnolaimus maximus*) life history: consequences for spawning biomass and egg production models. J. Northw. Atl. Fish. Sci., Vol. 41: 1–12
SEDAR37-RD-06	Munoz, R.C.; Burton, M.L.; Brennan, K.J.; Parker, R.O. Jr. 2010. Reproduction, habitat utilization, and movements of hogfish (*Lachnolaimus maximus*) in the Florida Keys, U.S.A.: comparisons from fished versus unfished habitats. Bulletin of Marine Science, 86(1): 93–116.

Appendix 2: Statement of Work

This appendix contains the Statement of Work, including two annexes, that formed part of the consulting agreement between Northern Taiga Ventures Inc. and the author. Note that the dates in the Tentative Schedule of Milestones and Deliverables therein were not revised when the timing of the review was changed.

External Independent Peer Review by the Center for Independent Experts

SEDAR 37: South Atlantic and Gulf of Mexico Hogfish Assessment Desk 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 37 will be a compilation of data, a benchmark assessment of the stock, and CIE assessment review conducted for South Atlantic and Gulf of Mexico hogfish. The desk review provides an independent peer review of SEDAR stock assessments. The review is responsible for ensuring that the best possible assessment is provided through the SEDAR process and will provide guidance to the SEFSC to aid in their review and determination of best available science, and when determining if the assessment is useful for management. The stocks assessed through SEDAR 37 are within the jurisdiction of the South Atlantic and Gulf of Mexico Fishery Management Councils, and the states of Texas, Louisiana, Mississippi, Alabama, Florida, Georgia, South Carolina, and North Carolina. 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 statement of work (SoW) tasks and terms of reference (ToRs) specified herein. The CIE reviewers shall have expertise in stock assessment, statistics, fisheries science, and marine biology sufficient to complete the tasks of the peer-review described herein. 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 participate and conduct an independent peer review as a desk review; therefore travel will not be required.

Statement of Tasks: Each CIE reviewer 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 contact information 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 assessment and other pertinent background documents for the peer review. 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 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 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 impartial and independent peer review in accordance with the tasks and ToRs specified herein, and each ToRs must be addressed (**Annex 2**).
3) No later than June 30, 2014, each CIE reviewer shall submit an independent peer review report addressed to the "Center for Independent Experts," and sent to Dr. 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**.

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

15 May 2014	CIE sends reviewer contact information to the COR, who then sends this to the NMFS Project Contact
1 June 2014	NMFS Project Contact sends the stock assessment report and background documents to the CIE reviewers.
9-20 June 2014	Each reviewer shall conduct an independent desk peer review
30 June 2014	CIE reviewers submit draft CIE independent peer review reports to the CIE Lead Coordinator and CIE Regional Coordinator
13 July 2014	CIE submits CIE independent peer review reports to the COR

20 July 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

Allen Shimada
NMFS Office of Science and Technology
1315 East West Hwy, SSMC3, F/ST4, Silver Spring, MD 20910
Allen.Shimada@noaa.gov Phone: 301-427-8174

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

SEPTEMBER 2014

Key Personnel:

NMFS Project Contact:

Julie Neer, SEDAR Coordinator
4055 Faber Place Drive, Suite 201
North Charleston, SC 29405
Julie.Neer@safmc.net Phone: 843-571-4366

Appendix 2, 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, regardless of whether or not they read the summary report. The CIE independent report shall be an independent peer review of each ToRs, and shall not simply repeat the contents of the summary report.

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 2, Annex 2 – Terms of Reference

SEDAR 37: South Atlantic and Gulf of Mexico Hogfish Assessment Desk Review

1. Evaluate the data used in the assessment, addressing the following:
   a) Are data decisions made by the assessment panel sound and robust?
   b) Are data uncertainties acknowledged, reported, and within normal or expected levels?
   c) Are data applied properly within the assessment model?
   d) 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:
   a) Are abundance, exploitation, and biomass estimates reliable, consistent with input data and population biological characteristics, and useful to support status inferences?
   b) Is the stock overfished? What information helps you reach this conclusion?
   c) Is the stock undergoing overfishing? What information helps you reach this conclusion?
   d) Is there an informative stock recruitment relationship? Is the stock recruitment curve reliable and useful for evaluation of productivity and future stock conditions?
   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?
4. Evaluate the stock projections, rebuilding timeframes, and generation times, addressing the following:
   a) Are the methods consistent with accepted practices and available data?
   b) Are the methods appropriate for the assessment model and outputs?
   c) Are the results informative and robust, and useful to support inferences of probable future conditions?
   d) 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 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.
8. Prepare a Peer Review Report summarizing the Reviewer's evaluation of the stock assessment and addressing each Term of Reference.