Response to SARC 55 Summary Report regarding Gulf of Maine Cod

Doug S. Butterworth and Rebecca A. Rademeyer

January 2013

Introduction

In offering this response, we appreciate (as the SARC Committee themselves acknowledge in their Report) that limitations of time precluded their full consideration of the evidence available. Naturally we must point out the problems we see with certain (and key) parts of this Report, but we acknowledge that at least some of those problems may be a consequence only of inadequate availability of time to clarify issues at the Committee's public sessions.

We respond under four headings. The material below intends a brief summary – the technical details of the arguments providing justification for our comments are to be found in the Appendix.

Natural Mortality

The SARC Committee's Report mischaracterises the main issue under this heading, which was not whether *M* has changed, but rather what the appropriate value of *M* is, given the first direct estimates from tagging studies becoming newly available, which their authors had argued to exclude values (for the period in the first decade of this century from which that data had been obtained) of less than about 0.4.

Furthermore the SARC Committee has erred in performing reference point computations for the *Mramp* scenario with M=0.2. The provisions of the National Standard Guidelines make quite clear that the value used should be 0.4, as corresponding to the currently prevailing conditions for that scenario.

Length of Time Series used in Assessment

The SARC Committee detail their concerns about assessments commencing earlier than 1982 in seven bulleted paragraphs. The arguments in each of these paragraphs are problematic for reasons that include errors of fact, apparent use of methodology known to be unreliable, lack of balance, inconsistencies and drawing inappropriate inferences.

Of greatest concern is an evident misunderstanding by the SARC Committee in respect of the primary reason that they give (and as they emphasised during plenary discussions) of "*concerns about the quality and the less detailed information available in the earlier part of the time series*". Wide experience with the use of such models in other fora has consistently indicated that estimates of quantities of importance for management purposes tend not to be very sensitive to such data features. Many sensitivity tests have been conducted to check this for the Gulf of Maine cod stock. Other scientists have frequently been requested to nominate further tests that might invalidate this general result in this case. In none of these cases has any appreciable sensitivity of management-related results been evident. Accordingly the SARC Committee's primary reason for concern falls away.

Multiple Models

The SARC Committee's comments here are predicated on a false assertion, as evidenced by their statement that:

"the SAW was unable to reach consensus on which model represented the best available science"

taken together with:

"because a consensus was not reached by the SAW Working Group, much more time was needed in the review to try to find a consensus (at least among the reviewers)"

As explained further in the Appendix, the implicit assumption being made here by the SARC Committee - that the best available science requires consensus on a single assessment - is quite incorrect. To the contrary, the most fundamental aspect of fisheries science, because of pervasive inherent uncertainties, is that many different assessments of a resource are defensible, and the challenge is how to take this range into account in providing advice compatible with use of the "best available scientific information" as required by US legislation. A single assessment, in the sense that the SARC Committee appear to use the term, will certainly not be able to effectively capture this range in all cases. This is why model averaging informed by risk analysis may need to be used, particularly in circumstances such as those which pertain to Gulf of Maine cod, as the best approximation available within the assessment paradigm to the Management Procedure (MSE) approach generally regarded as a superior basis to formulate scientific recommendations for management measures such as catch limits.

To that end, Table 1 is provided as an illustration of the range of assessment models over which some weighted average needs to be taken in developing a recommendation for a catch limit for Gulf of Maine cod.

In such an exercise, care must be taken to avoid a problem frequently evident in the SARC Committee's report: that of confounding unbiased selection of defensible assessment models with considerations of risk. While clearly a SARC Committee can and should offer comments in that last regard, their responsibility is not to make (implicit) choices, but to forward information which facilitates such choice to bodies such as the Council and SSC where such policy-related aspects are to be determined. An obvious problem which arises when assessment advice is confounded by incorporation of risk considerations is double counting: for example, the SSC advises on catch limits based on $0.75F_{MSY}$ rather than F_{MSY} to take risk/uncertainty into account, but what if the assessment advice has already incorporated (some of) that uncertainty?

Process

The spirit of the SARC Committee's recommendation that: "*some thought should be put into establishing protocols and mechanisms for facilitating consensuses in both stages of the process (that is the SAW and the SARC)*" is appropriate, necessary and welcome. As explained above however, attention is also required in regard to exactly what matter it is about which consensus is needed – this will not necessarily be a single assessment as the SARC Committee apparently understand the term.

Table 1: SCAA-based results including estimates of BRP's and of 2013 catch limits under $0.75F_{MSY}$ or a proxy therefor for a number of alternative assessments. BH signifies (internal) estimation of the parameters of a Beverton-Holt stock recruitment relationship. Biomass and catch units are '000 mt. For the results for $F_{30\%}$ and $F_{40\%}$ BRPs, the average recruitment over the period 1982-2011 is used; if instead the average over the 1963-2011 period is used, the results are very similar. For cases for which there are comparable ASAP runs, results differ slightly as the SCAA and ASAP formulations use different procedures for shrinking recent recruitment estimates to some mean for the recruitment estimates for recent years. Note that the -lnL values are not comparable across the alternative starting years for the assessments.

Natural mortality	*M* =0.2		*M* ramp			*M* =0.4	
Start year	1932	1982	1932	1932	1982	1932	1982
SR	BH	no SR	BH	BH, *h* =0.8	no SR	BH	no SR
-lnL: overall	-2745	*-2133*	-2751	-2750	*-2138*	-2752	*-2134*
B^{sp}_{2011}	14.38	12.49	12.74	12.63	12.94	18.62	15.47
K	193.02	-	33.97	36.79	-	89.51	-
h	0.92	-	0.98*	0.80	-	0.77	-
F_{MSY}	0.31	-	0.95	0.89	-	0.77	-
$F_{30\%}$	0.26	0.26	0.73	0.73	0.73	0.74	0.73
$F_{40\%}$	0.18	0.18	0.43	0.43	0.43	0.44	0.44
B^{sp}_{MSY}	46.31	-	8.57	7.96	-	20.78	-
$B^{sp}_{30\%}$	42.57	40.71	10.89	10.87	10.64	17.79	16.79
$B^{sp}_{40\%}$	56.75	54.28	14.52	14.49	14.19	23.72	22.39
$B^{sp}_{2011}/B^{sp}_{MSY}$	0.31	-	1.49	1.59	-	0.90	-
$B^{sp}_{2011}/B^{sp}_{30\%}$	0.34	0.31	1.17	1.16	1.22	1.05	0.92
$B^{sp}_{2011}/B^{sp}_{40\%}$	0.25	0.23	0.88	0.87	0.91	0.79	0.69
$C_{2013}(0.75F_{MSY})$	3.87	-	5.07	4.70	-	7.83	-
$C_{2013}(0.75F_{30\%})$	3.29	2.60	4.06	4.02	4.33	7.54	5.53
$C_{2013}(0.75F_{40\%})$	2.30	1.82	2.63	2.60	2.82	4.87	3.59

* Constraint bound

Appendix

At times it is convenient for the reader to present extracts from the SARC Summary Report when providing specific responses. Where this is done these extracts are shown in *italics*.

Natural Mortality

The SARC Committee characterise this issue as: "*Has natural mortality changed over the years? If so will it continue to change?*". They do however acknowledge that: "*In three days, the Review Panel did not have the time to sort through all the possible evidence supporting these issues and apparently the Working Group didn't either.*"

It is therefore perhaps understandable that their characterisation and subsequent discussion seem to have missed almost entirely the key importance of what was the central new piece of evidence available to this assessment that had not been available to earlier ones. This is the first direct estimation of *M* (pertaining to the first decade of the century) from tagging studies. These estimates were first introduced by their NEFSC authors at the preceding SAW Models Issues meeting in October with the commentary that no plausible variation of the models which they had investigated was compatible with *M* estimates of less than about 0.4. The only other estimate of *M* cited in the "Cons" list of the following SAW Modeling and Reference Points meeting is the *M*=0.2 suggested by a meta-analysis of life history relationships. This is weak evidence because many of these relationships are highly correlated with each other, and they provide estimates which are very imprecise. Thus the main issue is the value of *M*, with the core evidence favouring the higher value (incidentally supported also by the model fits – see – lnL values in Table 1). Weaknesses in the evidence available for any change over time need to be interpreted in this context. Absence of quite plausible models reported to the SARC with *M*=0.4 (constant) are a reflection of the lack of time in the SAW Working Groups to which the SARC Committee makes appropriate reference; the need to progress through a lengthy agenda in two successive meetings precluded a return to reconsider the relative merits of *M*=0.4 *vs Mramp*, which quite likely would have resulted in tabling results for the former as well. (For this reason, such results have also been included in Table 1.)

We consider the SARC Committee's decision to move forward with assessments with more than one assumption for *M* entirely appropriate (see also the **Multiple Models** section below). However the Committee have clearly erred in recommending reference points for the *Mramp* scenario to be based on a value of *M*=0.2 rather than the value for the present time under that scenario of *M*=0.4. There are two reasons which justify this statement:

i) Though perhaps not entirely compelling, there is the information in the preceding paragraph which indicates that a value of *M* greater than 0.2 is more strongly favoured by the information available.

ii) Completely compelling though is information brought to our attention after the SARC meeting. The National Standard Guidelines for the Magnuson-Stevens Act, in terms of which advice from these assessments is being developed, state that MSY (and hence its associated reference points) is to pertain to "prevailing ecological, environmental conditions". (Incidentally a NMFS scientist checking the code for our SCAA model corrected us similarly: that in calculating reference points we needed to use weight-at-age vectors for the recent period rather than the year in which our analysis commenced.) Clearly then estimates from the tagging data which apply to the recent past are what are to be used, which is consistent with the basis for choosing the more recent value for *M* under *Mramp*, i.e. *M*=0.4. Aside from self-evident reasons of balance which provide important justification for this specification, it avoids the difficulty in which the SARC Committee found itself in having to consider how soon in the future prevailing conditions might change.

The Reference point values given for the *Mramp* scenario in Table 1 are based on the value of M=0.4 to correct for this error by the SARC Committee.

Length of Time Series used in Assessments

The SARC Committee's reasons for rejecting assessments starting earlier than 1982 are given in seven bulleted points.

I) The F_{MSY} reference point derived from the Ricker model based on the longer data series was sometimes higher than total mortality derived from surveys suggesting that F_{MSY} estimated in this way is higher than would make sense as the stock decreased at these mortality levels. The Review Panel acknowledges that the criterion for determining survey total mortality integrates selectivity as well, but believes the above argument still holds.

Note that this links also to a "Con" comment recorded in the preceding SAW Modeling and Reference Points meeting:

From the 1970s forward Fs and catches consistent with Ricker-based F_{MSY} caused SSB declines

Fig. A1 shows the results from a run of the SCAA model with a Ricker stock-recruitment relationship and M=0.2 (the *Mramp* scenario gives qualitatively similar results). There are three distinct periods in terms of the (fully selected) F value compared to F_{MSY}.

- Prior to 1982: F is initially low and later increases. Spawning biomass is generally above B_{MSY}, first increasing, but then declining as to be expected for a biomass above B_{MSY} when F approaches F_{MSY}. (There is some variation about the trend as a consequence of recruitments less than predicted by the Ricker relationship during the late 1960s.)
- From 1982 to 1998: F is above F_{MSY}, and by quite an extent in the 1990s, causing spawning biomass to drop well below B_{MSY}.
- From 1998 onwards: F fluctuates close to F_{MSY}, with spawning biomass showing a broadly increasing trend towards B_{MSY} as would be expected. (Again superimposed on this trend are the consequences of some poor recruitments compared to the Ricker predictions across the turn of the century.)

Thus this particular assessment is entirely self-consistent, as to be expected from a model fitting the data without evidence of any serious model mis-specifications. The "Con" above and seemingly also the first sentence of the quote from the SARC Committee's Report are incorrect – they presumably arose from a failure to recognise that for $F= F_{MSY}$, one would expect spawning biomass to decline in periods where it was in excess of B_{MSY}. The SARC Committee's Report does not make entirely clear how they estimated "total mortality estimated from surveys", though one presumes this was from some type of catch curve analysis. Such approaches are known to give unreliable results, as well demonstrated by work over 20 years ago in the International Whaling Commission's Scientific Committee, which consequently no longer uses that approach (see for example Butterworth and Punt, 1990, and references therein). The information from age data in surveys, as reflected for example by the slope of a catch curve, is a combination of the effects of recruitment trends, present fishing mortality, the trend in selectivity-at-age, natural mortality and an integral of the past effects of fishing. In the absence of other information, which requires a full assessment model to be taken into account properly, these different effects are confounded and, except in special circumstances whose applicability would first need to be justified (though hardly seem likely to apply in this case given

relatively substantial and time-varying catches), such approaches will not yield reliable estimates, thus precluding drawing inferences on this basis as the SARC Committee has apparently done.

II) Although the Ricker model fit the longer data series better than other models (neither the Ricker or Beverton-Holt could be reasonably fit without including some other information, as that derived from the longer data series or some other external piece of prior information), the fit was clearly influenced by low recruitments in earlier years associated with high spawning stock biomass (SSB). The Review Panel could not decide if this was a period with low recruitment productivity driven by external forces or if it was a low recruitment period because of high SSB.

The reason given by the SARC Committee does not exclude the possibility of a Ricker stock recruitment curve, so that in terms of defensible models it remains on the table. One can always readily speculate about other effects than spawning biomass driving recruitment patterns. For example, the recruitments of the early and mid 1980s are relatively high, and are particularly influential in the conclusion drawn from the ASAP model starting in 1982 (which was favoured by the SARC Committee) that the stock is presently overfished; why can one not equally defensibly argue that these 1980s recruitments were driven by external forces and should not be included in the assessment or taken into account in estimating reference point values? This is why it is customary to require statistical tests, such as use of the STARS method (Rodionov, 2004), before accepting arguments (essentially of regime shifts) of the nature which the SARC Committee advances here as defensible alternatives (let alone preferred alternatives) to the assumption of stationarity.

III) The Beverton-Holt stock-recruitment model was similarly rejected because these low recruitment points also inflated the steepness parameter to values beyond what seemed reasonable.

For the *Mramp* scenario the steepness estimate hits an upper constraint boundary of 0.98 which indeed is defensibly rejected as unreasonable. In essence the SARC Committee is saying here that they have a prior for steepness, presumably based on data for some other stocks, which excludes that 0.98 value at least, and that is certainly a reasonable and defensible position to take. But if for that *Mramp* scenario steepness is fixed at 0.8, which is certainly a value perfectly compatible with assessments of many other stocks, the key management quantity output (the catch limit for 2013) is little altered (by only some 7% - see Table 1), and the argument given above provides no basis to reject that result. Estimates for steepness for other *M* scenarios in Table 1 do not seem unreasonable, but if the SARC Committee does consider them to be outside the bounds of their steepness prior, their appropriate response would be to specify that prior so that computations could be repeated in the same way as in Table 1 for the *Mramp* case. Already the results for that last case indicate that the resultant 2013 catch limit output would also not be greatly affected for the other scenarios for *M*.

IV) Including the earlier catch series was necessary to fit a stock recruit relationship, however, because of the above arguments and concerns about the quality and the less detailed information available in earlier part of the data series, the Review Panel concluded that these relationships were too unreliable to provide MSY reference points for characterizing assessment advice and so all model formulations (either ASAP and SCAA) that included a stock recruitment relationship were not considered further.

Responses have already been provided to the "above arguments". It is important to address the comment about "*concerns about the quality and the less detailed information available in the earlier*

part of the data series", particularly as the Chair of the SARC Committee when questioned twice on the reason for not including a longer timeframe assessment in the Committee's recommendations, replied on both occasions that it was because of comments from the audience (referring to NEFSC staff) that the extra data in the longer timeframe were too unreliable.

From discussions that took place during the SARC meeting, we can identify only three sources of comments that the SARC Committee could have construed as reflecting unreliability of these earlier data.

- As noted above, a result of assessments making use of these earlier data was that Gulf of Maine cod recruitment in the 1960s was generally low. One member of the audience, drawing attention to the value of only one datum, argued that there was no basis to conclude from the data available that this was the case. This despite the assessment, which is an integration of all the available data to draw statistical inferences, showing these recruitments were low and statistically quite precisely determined. Nevertheless the "reason" for this result was requested by this audience member. A simple approach to demonstrate the reason was suggested, and endorsed by Dr. Methot, but it required an overnight computer run. However the Chair ruled that because of shortage of time, it would not be possible to consider such results the next day. The Chair's difficult position is understood, though it should be noted that the approach suggested was nevertheless run overnight. The results of that run are provided in Fig. A2, which shows the consequences of forcing recruitments in the 1960s to be equal to the average of recruitment levels over the following 14 year period in line with the alternative argued by that audience member. There is a substantial deterioration in the negative log likelihood of over 100 points, primarily a consequence of serious deterioration in the fits to the age-proportion data from the NEFSC surveys and to the annual catches. The latter is required to climb to a value in the early 1970s which is more than double that recorded for any of the past 80 years. There is therefore clearly no substance in the assertion made by this member of the audience. It should also be noted that an alternative explanation for the ability of the data to provide a good determination of these 1960s recruitment estimates had already been provided in the paper we submitted to the March 2012 SSC meeting (Butterworth and Rademeyer, 2012). This issue was raised in the preceding SAW Modeling and Reference Points meeting and that explanation was again presented. No reservations about that explanation are registered in the "Cons" comments recorded in the report of that meeting, presumably because all regular participants in the meeting were satisfied that the question raised had been "asked and answered".
- Another audience member raised the matter of analyses pointing to loss over time of some local spawning aggregations of the Gulf of Maine cod stock, specifically that the chronology of local depletions documented by Ames (2004) should be considered, presumably also in the decision about what year to start the assessment. However, the detail of the paper was not discussed during the SARC meeting. According to Ames (2004), spawning components were lost principally sometime between the 1920s and late 1940s. Therefore, the critical issue of recruitment during the 1960s relates to a period after the loss of spawning components, thereby invalidating any relevance of the concern stated to the key point at issue during the SARC meeting. Furthermore the October SAW workshops had considered such matters and decided nevertheless to go forward on the basis of a single stock assumption for the assessment.
- Most importantly though, wide experience in other fora with the use of models such as SCAA to cover longish time periods has shown that estimates of key quantities required for management advice are relatively insensitive to certain uncertainties associated with, and even absence of, some earlier data. There are certainly uncertainties in the data for Gulf of Maine cod for the 60s and 70s which are greater than those for the following decades. Despite this experience from other applications, it remains reasonable to require sensitivity analyses to check that these customary results remain valid given the particular uncertainties that apply in

this specific cod case. Sensitivity computations had indeed been pursued to confirm this. It is conceivable that the SARC Committee understood comments from some audience members to imply that this sensitivity testing had been inadequate. This hardly seems a supportable conclusion given the process that was followed. NEFSC staff were asked on many occasions to nominate alternative assumptions concerning these uncertainties in formulating sensitivity checks: specifically at the October 2011 SAW assessment meeting, in the first draft (available in November 2011) of the paper eventually submitted to the March 2012 SSC meeting (Butterworth and Rademeyer, 2012), at that SSC meeting itself, in subsequent email correspondence (to which Dr Palmer responded constructively, with his suggestions being taken on board, though the consequent sensitivity runs showed sensitivity to the factor he raised to be minimal), and at the October 2012 SAW workshops. At those last workshops (quite large) variances were specified for historic catch estimates, though subsequent analyses again evidenced the customary pattern that key management-related results were not greatly affected. Analyses were presented at those workshops by NEFSC staff which showed that the catch adjustments to include recreational fishing and discard mortality might vary with time rather than be constant over time as assumed. We asked the question at the SAW Modeling and Reference Points meeting whether there was a need to explore further sensitivity runs in relation to this effect, but the response given was that this was not necessary.

Given the above, we are unable to identify any cogent evidence to support the SARC Committee's contention that "*concerns about the quality and the less detailed information available in the earlier part of the data series*" are sufficient to invalidate estimates of quantities of importance for management from assessments which include those earlier data.

V) *Regarding the low recruitment values of the 1960s, it looked like there were other avenues that could be pursued to help validate whether or not they should be included in determining stock recruitment model fits and associated reference point calculations. For example, examining evidence of ecosystem drivers would help determine if these recruitments were more likely to be evidence of density dependence or alternatively an environmental regime shift or a change in predation by other species. A general concern about the quality of the data in the earlier part of the series provides further motivation for examining the credibility of these influential points.*

The response under *II)* applies also here. While appreciating the constructive intent of the suggestions made, the probability that they might yield reliable results can hardly be considered as other than rather small when account is taken of the great volume of evidence in the literature of later failure of proposed environment-recruitment relationships (see summary in Myers, 1998), and the very small number of fisheries where such relationships are actually used explicitly in the determination of quantitative management advice. Indeed, the classic and frequently cited example of such a case – the use of temperature data in determining catch limits for Californian sardine – has recently had to be added to the casualty list (McClatchie *et al*., 2010).

VI) As no standard stock-recruitment relationship could be found, the use of proxy reference points for this stock was supported.

The response under *III)* above also applies here. The SARC Committee may have a prior that precludes consideration of domed relationships such as Ricker. But their comment *III)* above indicates that they find the Beverton-Holt relationship acceptable provided the steepness estimate is not too high. What then necessarily excludes consideration of results from assessments incorporating their (implicit) prior on the steepness parameter, with this in turn providing the SARC Committee a basis to obtain a defensible direct estimate of F_{MSY}?

There is also an inconsistency in the SARC Committee's conclusion to accordingly use proxy reference points when their ultimate recommendation is use of the F40% proxy. This particular reference point has its origins in analyses based on the use of a range of both Beverton-Holt and Ricker stock recruitment relationship forms (Clark, 1991). If for the Gulf of Maine cod stock, the range considered by Clark can be narrowed given the information available for that stock, and indeed even the Beverton-Holt forms estimated when including the earlier data have reasonable associated precision, how can what is nothing other than a narrowing of the range considered by Clark through use of the data for the stock concerned not be classified as Best Available Scientific Information in preference to use of a generically based proxy?

VII) One other important related issue should be noted when using the Ricker or the Beverton-Holt relationships for data like these. The two models result in very different SSB_{MSY} and F_{MSY} reference points although the resulting recruitment levels at these points may be close to indistinguishable. Basing overfishing thresholds on such a volatile criterion may not be the best approach for establishing stable and sustainable management actions for stocks with this type of recruitment history.

The response under *VI)* also applies here. The pertinence of the observation that R_{MSY} estimates may be very similar for the two models is unclear – self-evidently, if both estimates of SSB_{MSY} fall within the range of the data so that extrapolation is not involved, this is likely to be the case anyway. Indeed, if anything, this observation constitutes an argument for preferring the Ricker results, in that the shape of estimated Beverton-Holt curve is such that the gain in sustainable yield in increasing from the Ricker estimate of SSB_{MSY} to the Beverton-Holt estimate (if the latter is a more accurate reflection of the underlying reality) is so small that the loss of catch in the short to medium term to achieve that further SSB increase would be entirely unjustifiable under any socio-economic analysis, even under circumstances where the Ricker result was accorded a rather lesser weight than the Beverton-Holt for some reason. Furthermore, if it is the Ricker curve that does indeed reflect this reality, excluding this from consideration and rebuilding the resource to the level suggested by the estimated Beverton-Holt curve would see the resource at an abundance where the only available information indicates recruitment (and consequently yield) to be low, which again would hardly be defensible socio-economically.

In any case, though, this argument also fails to take into account the statistically appropriate weighting to be accorded to the two models. It would have validity if this weighting was near equal, and decisions were based on choosing the one with the greater likelihood. In such cases clearly values of reference points could swing wildly and undesirably from one assessment to the next as more data became available. But in this case, unless one imposes a prior that excludes Ricker-like forms, for the M=0.2 scenario for example, the Ricker variant is preferred to the Beverton Holt by 2.8 log likelihood points, i.e. if AIC weighting were used, the relative probabilities to be accorded to the two models is about 94% to 6%, so that the volatility to be expected under near equal probabilities is unlikely to occur in this instance.

The SARC Committee is correct to be concerned about high temporal variability of estimates. But that is precisely why appropriately weighted model averaging approaches are the correct management approach in circumstances of alternative defensible assessment models. Thus the appropriate inferences to be drawn from this observation by the SARC Committee are actually the reverse of some of the key conclusions which they reach.

Multiple Models

The tabling of alternative assessment models obviously created a challenge for the SARC Committee. The preceding October SAW working group meetings had already partly advanced how this could be addressed by initiating risk analysis computations for these models (incidentally an approach suggested at the March 2012 SSC meeting). Those meetings however ended before advancing to the standard final stage of this approach of assigning weights and taking appropriate averages over the range of defensible models presented.

However, when at the SARC meeting one of us (DSB) proposed this way to proceed and resolve the matter, the Chair immediately ruled it out of consideration, stating (presumably and perhaps understandably for the perceived reason of insufficient time) that "we're not getting into that". It is our contention that in so doing the Chair (unintentionally) may have precluded the achievement of recommendations based on the "Best Available Scientific Information" (BASI) required under the Magnuson-Stevens Act.

The SARC Committee report states in regard to such an averaging approach:

One alternative to such an approach, used in other parts of the world, is model averaging, whereby the results of alternative models are averaged for the purposes of developing management actions. ……. This part of the process can be undercut if the option of model averaging is relied on too quickly. Furthermore, model averaging can sometimes hide the consequences associated with alternative assumptions about the state of the system. Consequently, here we put forward several alternative assessments, but would suggest that they be viewed separately in terms of their assumptions, outcomes, and consequences rather than being averaged for decision making

Importantly this text reflects internal SARC Committee discussions **after** the plenary sessions. The only comments made during those sessions are as reflected above. The proposal which the Chair precluded being developed at the meeting was **not** of the form of the extreme of simple quick averaging in isolation which is correctly condemned above, but rather one, as always should be the case, of factoring in risk analysis considerations. Thus, for example, while given the data it is quite inappropriate to exclude an assessment based on the Ricker stock-recruitment relationship from consideration in an unbiased process based on BASI, it might well be defensible to give it a weight amounting to effective exclusion when risk considerations are taken into account in the averaging process.

The practice "*to put forward a preferred assessment among the multiple available that best characterizes the state of the system*" is self-evidently compatible with BASI **only** under certain circumstances, which are when that assessment is either unquestionably better than the alternatives or is (implicitly) agreed to reflect an unbiased proxy for an appropriate weighted average which includes those alternatives. This is because of universal recognition (because fisheries is an inexact science) that there are many alternative plausible/defensible assessments possible for any stock – failure to properly take account of this range (for example by excluding assessments based on longer time series of data) cannot be compatible with BASI because it is ignoring pertinent information.

The SARC Committee's concern about complications of alternative reference points in the situation of multiple models:

We recognize that this will likely complicate the management process by requiring multiple reference points, possibly alternative conclusions about stock status, and multiple methods for deriving stock projections.

falls away given the use of model averaging. Each model returns not only estimates of, for example, reference points, but also associated estimates of precision – effectively distributions. On weighted averaging, these distributions sum to a composite distribution for each quantity, and these readily provide single values for the quantities required through use of an appropriate distribution summary statistic such as the median. This averaging really reflects the development of a single (though composite) model – in essence it is taking the standard Bayesian assessment approach one stage further.

The incompatibility with BASI of a use of only the preferred model choice amongst many defensible models in a situation such as applies for Gulf of Maine cod is readily demonstrated by considering the implications of comments by SARC Committee members in their discussions towards the end of the SARC meeting plenary. It was evident that they were struggling to find any basis in the information available to provide a preference between M=0.2 and *Mramp*, and between F40% and F30% as an MSY proxy. Consider the associated possible catch limit recommendations for 2013 for the consequent four possible scenarios for the assessments starting in 1982 with no stock-recruitment relationship, whose results are shown in Table 1:

M=0.2	F40%:	1.82 kt
Mramp	F40%:	2.82 kt
M=0.2	F30%:	2.60 kt
Mramp	F30%:	4.33 kt

Say the SARC Committee had eventually characterised their views as a 51% preference for M=0.2 and for F40%. Then under the "*single preferred assessment*" approach, their advice is for a catch limit of 1.82 kt. But this is in the lower tail of the overall distribution that reflects their views. It does not provide unbiased advice as required under BASI because effectively it is ignoring pertinent information. Furthermore, given that a subsequent SARC Committee could readily amend these preference weightings each from say 51% to 49%, leading them to recommend a catch limit of 4.33 kt, the approach leads inevitably to the very volatility of decisions that the SARC Committee correctly seeks to avoid (see *VII)* in the section above). To put this in Bayesian terms, use of posterior medians is a much more robust and sensible basis for decision making than use of posterior modes.

Accordingly, consistent with the BASI requirement, final advice in the Gulf of Maine cod situation of near equally defensible models and related assumptions cannot be other than under some appropriate model averaging approach.

References

Ames EP. 2004. The stock structure of Atlantic cod in the Gulf of Maine. *Fisheries*, 29:10-28

Butterworth DS and Punt AE. 1990. Some preliminary examinations of the potential information content of age-structure data from Antarctic minke whale research catches. *Reports of the. International. Whaling. Commission* 40: 301-315.

Butterworth DS and Rademeyer RA. 2012. An investigation of differences amongst SCAA and ASAP assessment (including reference point) estimates for Gulf of Maine Cod. Document submitted to the March 2012 meeting of the New England Fisheries Management Council SSC. 34pp

Clark WG. 1991. Groundfish exploitation rates based on life history parameters. *Canadian Journal of Fisheries and Aquatic Sciences*, 48: 734-750.

McClatchie S, Goericke R. Auad G. and Hill K. 2010. Re-assessment of the stock–recruit and temperature–recruit relationships for Pacific sardine (*Sardinops sagax*). *Canadian Journal of Fisheries and Aquatic Sciences*, 67: 1782–1790.

Myers RAM. 1998. When do environment-recruitment correlations work? *Reviews in Fish Biology and Fisheries*, 8: 285-305.

Rodionov S. 2004. A sequential algorithm for testing climate regime shifts. *Geophysical Research Letters*, 31: L09204.

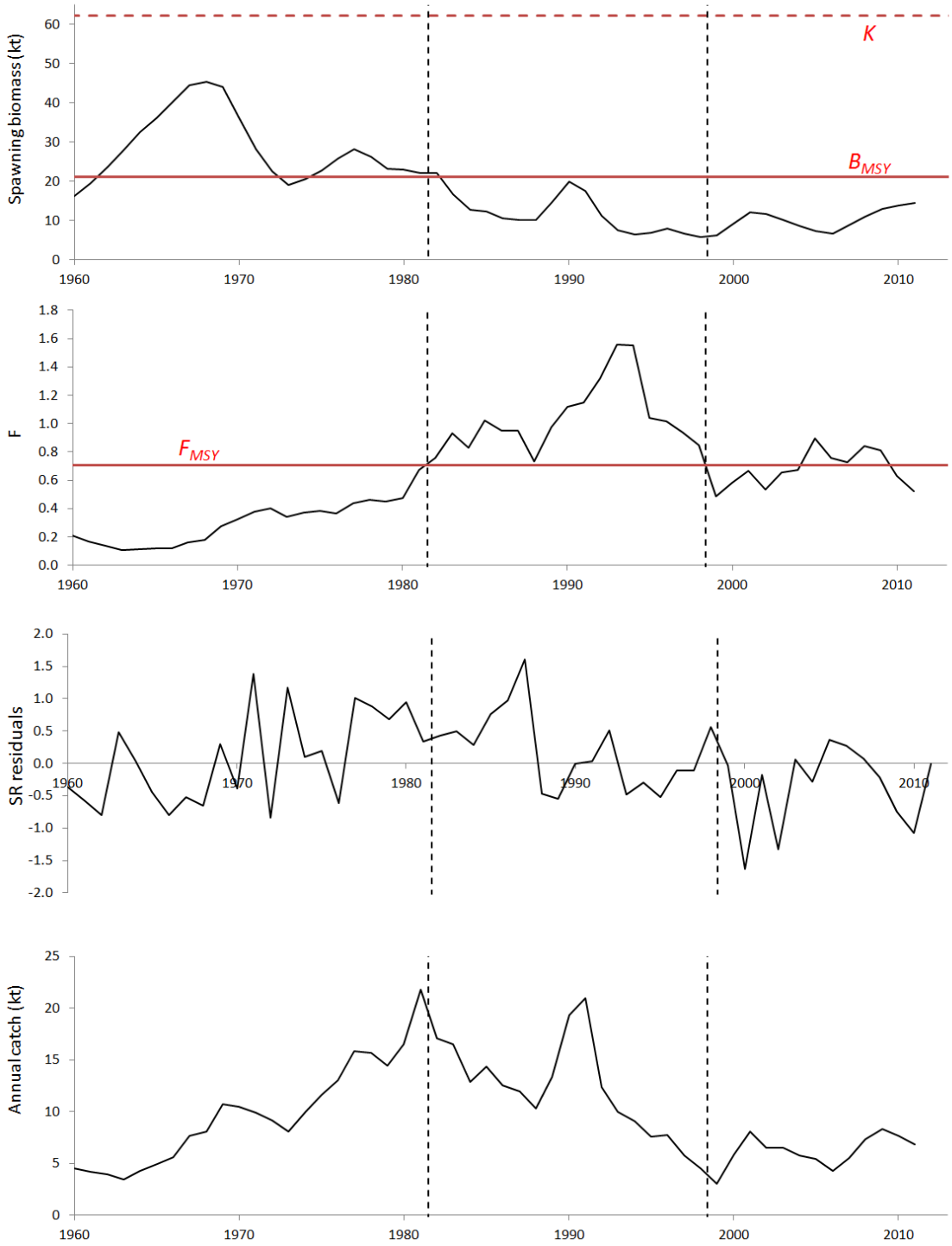

Fig. A1: Time trajectories of spawning biomass, fully selected fishing mortality, stock-recruit residuals and catches for the SCAA model with a Ricker stock-recruitment relationship, starting in 1932 and M=0.2.

	Start year 1932, BH, M =0.2		
	Standard	Fixed 60s R	Difference
-lnL: overall	-2745.1	-2620.3	124.8
-lnL: survey	-24.1	-18.4	5.7
-lnL: comCAA	-786.6	-786.5	0.1
-lnL: survCAA	-1812.8	-1746.5	66.3
-lnL: survCAL	-160.2	-148.9	11.3
-lnL: RecRes	35.6	27.6	-8.0
-lnL: Catch	3.0	52.4	49.4

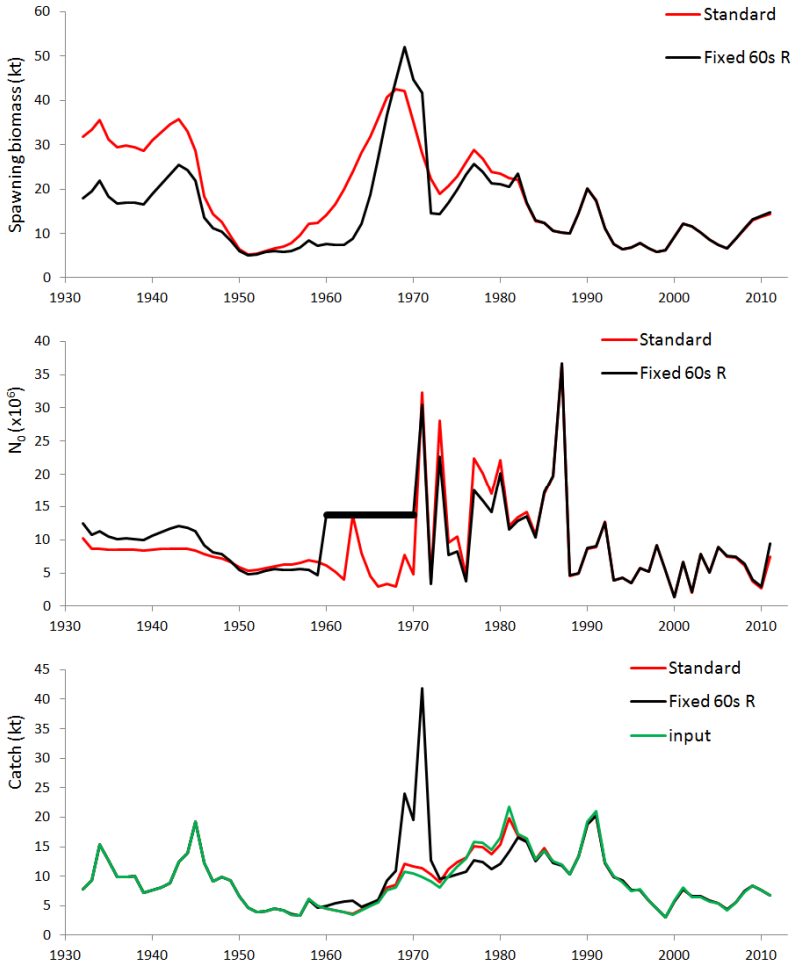

Fig. A2: A comparison of results for the "Standard" SCAA model fit starting in 1932 with M=0.2 and with a Beverton-Holt stock recruitment curve, and an alternative "Fixed 60s R" which fixes recruitments over the 1960s (as shown by the **bold** line segment in the central plot) to their average over the period 1971 to 1984. The table compares the contributions of various components to the overall negative log likelihood -lnL.