fit by the model (Figure 5). Of potential stock concern is that the NMFS longline survey is generally under the expected survey abundance since 2010 (Figure 5), suggesting that information on larger fish in the population added by this survey leads to a more pessimistic assessment of overall stock depletion (as indeed shown by model 15.6A results). However, the model is not fully tuned, so such supposition may be premature. However, it does highlight that if the index is to be used, some evaluation of possible bias in relation to Pacific cod, perhaps most importantly since 2010 is required. The model that includes the NMFS longline survey is able to fit the associated length compositions well.

Before deciding to include the NMFS longline index and associated lengths in a proposed central EBS SS model, an investigation into the properties of the EBS NMFS longline index in relation to Pacific cod in particular should be done. The investigation should particularly examine possible bias in the index since 2010 as this appears to be influential on assessment results. If the index is found unlikely to be biased, then I recommend inclusion in the model with additional sd estimated.

Aleutian Islands

The overall fits by the AI model to lengths (Figure 6) and the abundance index appear reasonable. Abundance index point estimates for 2004 and 2014 appear to most conflict with other information in the AI model.

Before deciding to include the NMFS longline index and associated lengths in a proposed central AI SS model, an investigation into the properties of the AI NMFS longline index in relation to Pacific cod in particular should be done. If the index is found unlikely to be biased, then I recommend inclusion in the model with additional sd estimated.

How should the various data sets be weighted?

For abundance index data, iterative reweighting to potentially allow additional index error was previously an accepted procedure for many US and Australian stock synthesis assessments. Such iteration was done manually, and more recently the ability to internally estimate additional index error (via an additional sd) has been added as an option to SS. Use of that option has become accepted practice for many recent assessments. Estimation of additional index error is normally done for all indices included in a stock assessment as (perhaps in my naive interpretation), the input variability usually only accounts for measurement error and the process error component is unknown.

Input sample sizes for composition data have an influence on assessment results and it has also become generally accepted practice for those sample sizes to more reflect the number of sampled fishing trips, rather than the number of fish measured.

Relative data weighting in stock assessments for composition data and the goal of standardized approaches has been the subject of recent and ongoing research particularly in the US west-coast, and the subject of a Center for the Advancement of Population Assessment Methodology (CAPAM) workshop in La Jolla, CA in October of 2015 (http://www.capamresearch.org/data-weighting/workshop). While there has been some recent narrowing down of agreed procedures among US west-coast stock assessors, it has also been recognized that it is not currently possible to recommend default procedures for composition and conditional age-at-length (CAAL) data. There is agreement that the Francis weighting approach is more appropriate in cases where the model is not correctly specified as it takes autocorrelation among composition data into account. It is also agreed that for a correctly specified model, the McAllister-Ianelli harmonic mean weighting method works well. Both of these procedures have been extended from marginal length or age composition data to conditional age-at-length (Francis A and B methods are available for CAAL, with Francis B potentially preferred). A possible further development that may provide a direction forward is using the Dirichlet multinomial likelihood (Thorson, 2014), although this method will require review and implementation in SS before it may be used. Recent simulation work has shown that the McAllister-Ianelli arithmetic mean procedure is inferior to other methods (Punt, In press).

<u>What form (i.e., Stock Synthesis "pattern") should be used for the selectivity functions?</u>

SS provides a large number of selectivity pattern options (14 size and 12 age patterns excluding special, discontinued and mirror – SS user manual v 3.24s). By far the most commonly used patterns in recent stock assessments are logistic for simple asymptotic selectivity or the double-normal (most often size pattern 24 or age pattern 20) where selectivity is allowed to be dome-shaped. The flexibility of the double-normal is usually sufficient to account for the wide range of single-peaked shapes that may be expected from a single fishing gear type. It is also possible to combine size and age selectivity patterns for a fishery or survey and to have differential selectivity by sex to, for example, account for reduced availability of older females in the population. To most easily account for "odd-shaped" selection that may be due to, for example, a combined fishery composed of several gear types, SS provides an age based selection pattern that generates an age-based random walk (age pattern 17).

Normally, fishery and survey selection is assumed to be primarily a length-based process as fishing gear selection is usually size-dependent. However, selectivity in an assessment model combines gear vulnerability with availability. Whether availability (e.g. due to migration, aggregation [e.g. for spawning], schooling) is age- or length-based is a more difficult question, so although length-based selection may be preferred for modeling, a case can still be made for age-based selectivity.

Generally, the selectivity pattern should be chosen (most likely from the options above) that has the fewest parameters, and allows an acceptable fit to the available composition data (e.g. no bands at particular lengths of significant length composition residuals). As surveys are designed to at least use the same fishing gear throughout, a good reason to use more complex patterns than logistic or double-normal would be required for those. If a fishery has fairly homogenous gear, a similar argument applies there as well. In the case of a fishery with mixed gear types, an opportunity exists to use a less restricted pattern shape, as provided by the age-based random walk. At present, I don't think a random-walk length-based pattern is available, so selectivity in that case is restricted to being age-based.

<u>Should the models be structured with respect to season?</u>

It is usual practice for SS models to separate input data from surveys and fisheries that have demonstrably different selectivity if data are available to do so. Normally, the minimum requirement to allow data partitioning according to season, gear type or area is that a number of years of length or age composition data that are believed to be representatively sampled are available within each partition. Partitioning of composition data is only usually necessary if summary length/age compositions from comparable partitions show obvious apparent selectivity differences. Partitioning may also be required for abundance indices if different trends are observed by partition.

Models that specifically address the exploration of alternative structures regarding selectivity partitions have been developed and were presented for the EBS, so the discussion here will be confined to models from that region.

Simple examination of aggregated length data for the EBS shelf trawl survey, the slope survey, longline fishery and NMFS and IPHC longline surveys (Figure 1) show a marked difference in the shelf trawl survey to all of the others. Unfortunately, the trawl and pot fisheries were not included, but we know from diagnostic output from model 11.5 that trawl fishery selectivity seems to be intermediate between the trawl survey and longline fishery, and the pot fishery seems similar to the longline fishery (Figure 8).

Also notable is that the Jan-Apr trawl fishery lengths show a peak that is consistent with longline fisheries during that period only, which corresponds to the spawning season. Conjecture has been made about possible movement of larger fish from the NBS area, although another explanation may be the movement of larger fish from waters targeted by the longline and pot fisheries into shelf trawl areas during the spawning season. There is little information available from tagging and none that can address the question of movement in and out of the NBS. The shelf trawl survey is made outside of the spawning season, and at that time, less of the larger fish seem to be available on the shelf, although tagging of a small number of fish does indicate apparent random movement of fish over the shelf during that time.

For modeling purposes, the model only requires that the composition of the fishery catches be adequately accounted for each year, and the more important population abundance trends are taken from surveys (at least for the models here). The difference in trawl fishery selection by season seems to be a feature that can be addressed through seasonal model structure. This is done to some extent with model 11.5, but the fit to the Jan-Apr trawl fishery length composition by that model is not particularly good (Figure 8). In addition to gear/season partitioning, a large number of time blocks that allow selectivity to vary through time have been used in model 11.5. It may be questioned whether such fine scale partitioning of the data are supportable if partitioning and blocking first needs to be justified depending on whether prior data examination or independent knowledge about changes in practices suggests that all of those partitions are necessary, and that sufficient data are available within each to allow estimation of a different selectivity pattern.

A new procedure for accounting for fishery selectivity has been proposed here in model 15.6 where an age-varying random walk is used to characterize the selectivity for all combined fisheries (trawl, longline and pots) each year. This procedure seems attractive given the high level of partitioning required for model 11.5. If such a procedure can provide a means of accounting for total fishery removals each year according to size/age, then it should be acceptable. Diagnostic plots for fishery lengths, both by year and combined for model 15.6, show rather good fits to available data (all residuals are also within the range -2.0 to 2.0). There is very little catch taken aged above about 8, so fixing selectivity above that age seems reasonable.

As the proportion of trawl catch to longline has changed considerably over time, it would be expected that large changes in the general pattern of selectivity would also be observed, that are somewhat evident in the plot (Figure 9), but of possible concern. Is the amount of change consistent with the broad movement of the fishery from trawl to longline over time?

Also of some concern is that the general fishery pattern for model 15.6 is dome-shaped, allowing the model some flexibility to generate cryptic spawning biomass. This is also an area of on-going work, and some diagnostics associated with it are in development or available from Github as additions to R4SS. At present, the available code only works for 2 sex models, so cannot be applied here, but could be further generalized to do so. The inclusion of surveys that are more directed towards the older fish in the population help to alleviate cryptic biomass problems, and is therefore a further reason to consider the addition of at least one longline survey to the base model.

I believe that options are only currently available in SS for a random walk by age for annual selectivity, as used for model 15.6. If the same was done by length, more parameters would be required (if 1cm size bins), or alternative bin patterns could be explored. Such a length-based exploration would be useful, should such capability be available in SS.

As many current SS assessments grapple with highly partitioned fishery data, such a procedure has the potential for resolving some of those problems also. I do not have previous personal experience with this procedure, and am reluctant to agree on its use without a supporting simulation study that confirms its equivalence or even superiority to a high degree of data partitioning. Such a study would be reasonably easy to design and carry out. However, I am willing to agree that it seems to provide a good resolution to the problem for the fishery selectivity in the EBS models.

Should the models be structured with respect to gear type?

As this question mostly relates to dealing with the fisheries and not surveys, the discussion under ToR 2.2.2c was generalized to address both season and gear type.

How much time variability should be allowed, and in which parameters?

The only population biological parameter allowed to vary with time in most SS stock assessments is annual recruitment levels. Cumulative information on annual recruitment strength is provided fairly directly by composition data, so the reasons especially for high peaks and troughs in recruitment are usually apparent in the available data. It has also been recognized that other parameters are likely to vary through time – in particular natural mortality, but also growth and maturity. For natural mortality it has been considered difficult to estimate time trends in changes without strong independent estimates for those changes, such as from ecosystem studies showing differences in predator abundance, and that time trends in M are difficult to disentangle from other factors such as catch mis-specification (e.g. see Brodziak et al., 2011). Allowing time variation in factors that directly affect productivity also lead to questions about choice of appropriate time periods for the selection of management reference points, and how to make appropriate stock projections.

Additional model parameters that may vary with time that are often dealt with using time-block methods are fishery/survey selectivity and catchability. As already mentioned, for fisheries that are not associated with an abundance index, a fairly freely estimated time-varying pattern (such as used for EBS model 15.6) may be acceptable if it suitably captures annual fishery removals by size/age. For surveys the situation differs. Surveys are the most important source of abundance information for the model, particularly because at least the gear selectivity can be maintained as a constant through time. Availability (either by age or year) is another matter, but is usually treated as a source of additional random error. If a true trend (or even a step) exists in either survey selectivity or catchability, then that survey is biased, and the bias needs to be accounted for, or the survey truncated, split or discarded. Such a bias would ideally be investigated and identified with a focused study and auxillary data not necessarily used in the assessment model. Adding annual time-variability to survey selectivity or catchability and finding that trends are estimated may simply be providing a means for the model to trade trends in population abundance to improve the fit to noisy composition data in preference to abundance indices. The reason that such a model might result in trends in survey selectivity or catchability are not readily apparent from standard input data sources, and may be difficult to diagnose. Results from estimation of annual variability for the EBS trawl survey catchability in model 15.6 (Figure 10) do exhibit some runs in residuals that may be of concern – particularly from 1993 to 1996. Time-changes in trawl survey selectivity as estimated by the EBS model 15.6 shows very little change through time, suggesting that time-variability in trawl survey selectivity as implemented is not required (Figure 11).

My own recommendation for now is that time variability should be allowed in a parameter when there is an available reliable data source that fairly directly measures such a change, and that a trend exists in that data source that needs to be captured by the assessment model. This situation only currently exists for recruitment and fishery selectivity in the EBS model. It also provides some support to consider time variability in weight-at-length or size-at-age if those data sets show considerable trends over time.

Others (e.g. Anders Nielsen, Jim Thorson) have proposed that a more appropriate way to deal with time variability is to use mixed-effects models with time-varying "nuisance" variables such as recruitment modeled as random effects. Improved solutions for time-varying parameters may be possible using all of the currently available data sources, if/when SS RE becomes available.

<u>What constraints, if any, should be placed on survey selectivity at older ages?</u>

The models examined during the review for the EBS seem to fairly clearly demonstrate that the trawl survey selectivity is dome-shaped. However, the possibility that the survey is in fact asymptotic has not been eliminated. The extent of the survey dome-shape may, for example, be confounded with M. It may be that different data sources are in conflict about the estimated value for M that can be diagnosed with a Piner profile plot of likelihood components. Exploration of age-specific M (e.g. starting with a Lorenzen function) could also be done.

A range of plausible alternative models should be explored, and the extent of the estimated dome selectivity for the trawl survey examined for each to see if the dome is consistently required. However, as the extent of the trawl survey dome is probably one of the major axes of uncertainty in the model at present, it should remain freely estimated and informed by the available data in any chosen base model, possibly with forcing more or less dome as sensitivity analyses in the final assessment.

<u>What constraints, if any, should be placed on survey catchability?</u>

Because of the history of the development and use of the trawl survey as an absolute index of abundance, there remains some belief that there is sufficient information available to determine at least a plausible acceptable range for survey q, and to some, that range could be perceived to be quite narrow. Much work has been directed towards net avoidance and how that might be compensated by a q adjustment. I believe that all major potential sources of error in survey q should at least be stated in an accessible document, and errors in those dimensions at least be qualitatively examined and ranked. Those should include avoidance and other gear-specific fish behavioral issues, and potential error in scaling the swept area estimates to the population using assumptions about the population distribution during the survey by depth and area, and also even the assumption of known stock boundaries. A qualitative evaluation such as this would probably make it clear that the true error in q is reasonably high. It would also assist to determine what priorities should be given to field studies that may be directed towards reduction of the error in survey q and adjustments required to scale area swept biomass estimates to the total (available given survey selectivity) population. An extension to a more quantitative evaluation of the potential errors may also lead to a prior distribution for EBS shelf bottom trawl survey q that can be generally agreed, and could then be used for modeling without much controversy. Without at least a comprehensive qualitative evaluation of all major error sources, decisions about rejection of models that estimate q based on how different the estimated q is from acceptable values remains difficult, and currently in the domain of pragmatic judgment.

I believe that models that estimate the shelf bottom trawl survey q using a fairly non-informative prior (as in model 15.6) should currently be preferred. Agreed bounds on prior survey q point estimates can be used as one of the acceptance criteria for particular models. I personally have a fairly high tolerance for those values (based however, on only a limited background knowledge for this particular survey), and am comfortable with at least a factor of 2.0 (0.5 – 2.0 times the initial point estimates).

Should additional surveys be added to the models, q values for Pacific cod for those are less well understood, and non-restrictive priors for those are preferable, with q estimated.

How should large gradients be dealt with in otherwise apparently converged models?

Large gradients are generally considered to be an indication of a problem. However, if the hessian can be inverted and jitters also indicate convergence, then perhaps the problem is only minor. I do not have any reason to doubt the explanation given in the EBS assessment document for why large gradients might occur, but it does suggest to me that the implementation of age selectivity pattern 17 requires a closer look to determine if the problem can be corrected (e.g. to determine whether it contains badly behaved/non-differentiable "if" statements).

Anything else on which the reviewers care to comment

Retrospectives

Diagnosis of retrospective bias in stock assessments has received considerable past attention in the literature and was also the subject of a BSAI/GOA working group in 2013 according to meeting background information. Despite this attention, research is on-going, and means for diagnosis and correction for retrospective patterns are not agreed. Several diagnostic measures are available including Mohn's ρ, the so-called Woods Hole ρ, and the RMSE method devised by the BSAI/GOA working group. I am familiar with two rules of thumb that can be used to diagnose retrospective patterns that need to be addressed in some way. The first and simplest is by Hurtado-Ferro et al. (2014) that says that "values of Mohn's ρ higher than 0.20 or lower than -0.15 for longer-lived species (upper and lower bounds of the 90% simulation intervals for the flatfish base case), or higher than 0.30 or lower than -0.22 for shorter-lived species (upper and lower bounds of the 90% simulation intervals for the sardine base case) should be cause for concern and taken as indicators of retrospective patterns." The second by Brooks and Legault (2015) from VPA assessments "is to plot the terminal year estimate of SSB(T) vs F(T) along with bootstrap percentiles and compare that to the point estimate when SSB(T) and F(T) are adjusted by ρSSB,7 and ρF,7, respectively" to see if the ρ-adjusted point estimate falls outside the bootstrap percentiles on either axis - see Brooks and Legault (2015) for details. Brooks and Legault (2015) also provide a procedure for adjustment of short-term projection results to account for substantial retrospective patterns. Ideally, the diagnostics for a model acceptable for use for management advice should not show significant retrospective bias. EBS model 11.5 and the initial AI SS models did show significant retrospective bias (at least according to the Hurtado-Ferro et al. (2014) rule of thumb) that indicated that results from those models are not reliable for use for management advice, and that improved alternative models should be sought, or at least a projection correction may be required. Further model explorations for both regions have found models that do not exhibit a strong retrospective bias, and on that basis would be judged as improved models. Retrospective bias provides evidence for model mis-specification, but of course, the lack of a retrospective bias does not prove that the model is correctly specified.

So-called Ianelli "squid plots" provide an additional useful means for looking at retrospective patterns in annual recruitment deviations, but have potential application to any parameter allowed to deviate annually in a model.

Catch uncertainty

As for many models, historical catch in particular is uncertain, and the best estimate of historical catch has been made using assumptions that seem supportable. However, the construction of alternative plausible historical catch scenarios would be useful for the determination of sensitivity of the model to that uncertainty.

Steepness

Tier 3 methods by default assume a steepness value of 1.0. A requested run using a steepness value of 0.7 shows that EBS results are somewhat sensitive to the choice of steepness value, and this dimension of uncertainty should be highlighted.

Regime change

A regime change in 1976-77 affecting log mean recruitment in EBS model 11.5 has been avoided in EBS model 15.6 by starting the latter model after the regime change. Shifts in 1989 and 1999 have also been suggested according to the ecosystem considerations in the assessment documentation. Regime change was not examined at all during the review, but is another potential source of model uncertainty.

Inclusion of marginal age composition vs CAAL data

At present, both the EBS and AI enter age-at-length data as marginal age distributions. There has been a gradual trend in stock assessments to make improved use of data from otoliths by entering the data into models as conditional age-at-length. During the review the general wisdom of this approach was questioned as it was mentioned that some recent assessments had reverted back to marginal age distributions. A standard approach for dealing with age-at-length data currently seems to be unavailable.

Reviewer 3: Jean-Jacques Maguire

Executive summary

From what was discussed during the meeting and the documentation reviewed, there are no objective reasons to reject the IPHC longline survey as an index of stock size, assuming it has been correctly put together and calculated. The IPHC longline survey data should be thoroughly investigated. It should be used in the assessment unless fatal flaws in the data, in the treatment of the data or in the survey methodology are identified. Similar to the IPHC longline survey, there are no objective reasons to reject the AFSC longline survey as an index of stock size. The AFSC longline survey should also be thoroughly investigated and used in the assessment unless fatal flaws in the data, in the data treatment or in the survey methodology are identified.

Regarding the form of the selectivity function, my preference would be to not allow too much flexibility in selectivity changes over time and to not allow strange patterns (e.g. figures 2.1.3 in the Eastern Bering Sea and 2A.11 and 2A.12 in the Aleutian Islands in the December 2015 SAFE report). If allowing these strange patterns is a condition of getting a good fit or convergence, this would be a sign that something else might be wrong. If allowed to change over time and age, the changes should be relatively smooth and not result in peculiar patterns. The reason(s) for the apparent differences in selectivity between the IPHC longline survey and the AFSC longline survey for lengths above 70cm should be further investigated.

It could be worth investigating further changes in growth (Figure 11), particularly with respect to the implications for the assessment as growth changes may have an influence on fishing mortality and population estimates.

In the Aleutian Islands area, it is unlikely that there is a single stock in the traditional understanding of the concept. Simpler form of monitoring and management, in close cooperation with the industry and possibly NGOs, could be a better way of protecting the resources and managing the fisheries.

One cannot model oneself out of lack of data, particularly for the Aleutian Islands assessment. Stock Synthesis has so much flexibility that, given sufficient time, a skilled user can probably get almost any stock trend from a dataset. Indices of abundance should be given more weight in the assessment than length composition. Age composition, particularly from the commercial fishery, but also from surveys or other indices of abundance can be very informative if analyzed appropriately. Information in the length composition is at best indirect information on changes in stock size.

Analytical retrospective analyses are routinely done for both stocks. Historical retrospective, where there are successive accepted assessments, is also informative and should be done to indicate how consistent the assessments have been over time. Simpler models, e.g. like Robin Cook's or surplus production models should be investigated. It is not necessary to go to Ensemble modeling, but looking at more than one modeling framework might be informative.

Should data from the IPHC longline survey be used in either assessment?

During the review meeting it was not clear if the raw data received from the International Pacific Halibut Commission (IPHC) had been treated appropriately to derive an index of stock size. Further work was conducted by the AFSC survey unit during the meeting, and it seems that the series shown in the excel spreadsheet "Survey index comparison (trawl surveys, longline surveys).xlsx" could be treated as an index of stock size. The appropriateness of the data and how it was treated to calculate an index should be further verified between now and the assessment meeting later in the year. What data were used and how they were used should also be documented.

The IPHC longline survey has been conducted every year since 1997. The survey covers the Eastern Bering Sea area well (Figure 1) but it is not clear if all stations were used in calculating a relative index or if only those on the slope were used.

The main index of abundance used in the Eastern Bering Sea stock assessment is the AFSC shelf trawl survey. The agreement between the AFSC shelf trawl survey and the IPHC longline survey is not very good (Figure 2). This could be due to different size selectivities and / or inherent variability in the data.

The IPHC longline survey sample larger individuals (Figure 3) and a lag between the two indices would therefore be expected. However, the sudden decreases in the relative index in 1999 and 2005, and similarly sudden increase in the following year are unlikely to reflect real changes in stock sizes. These anomalies warrant further investigations to try to identify what might cause them. If there are valid reasons to exclude those two points, it might be possible to reconcile the IPHC longline and AFSC shelf trawl survey time series taking into account that they sample different size groups.

Including longline surveys (IPHC or AFSC) in the assessment might alleviate concerns that the shelf trawl survey samples poorly larger sizes, either because large Pacific cod are outside of the surveyed area or because they are able to swim faster than the fishing gear and therefore escape capture. Including one or more indices of stock sizes for larger fish sizes therefore has the potential to improve the assessment and reduce the uncertainty in the population estimates of larger fish sizes.

Figure 4 shows the spawning stock biomass (SSB) trends for the Eastern Bering Sea Pacific cod stock from various model configurations. I have not been able to find a model where only the IPHC longline survey is added to the AFSC shelf trawl survey, but I think one was presented during the meeting. My memory is that including data from the IPHC longline survey in the assessment implies lower terminal year biomass than when only the shelf trawl survey is used. Figure 4, however, shows that when both the IPHC longline survey and the AFSC longline survey are added (model 15.6B), the SSB estimates are the lowest of the model considered. Adding the AFSC slope trawl survey (model 15.6C) implies essentially identical results to adding the two longline surveys. Adding only the AFSC longline survey (model 15.6A) results in a SSB trend that is markedly different from those of the other models considered.

From what was discussed during the meeting and the documentation reviewed, there are no objective reasons to reject the IPHC longline survey as an index of stock size, assuming it has been correctly put together and calculated. Its influence on the assessment results, however, when used along with the AFSC longline survey is puzzling.

The IPHC longline survey data should be thoroughly investigated. It should be used in the assessment unless fatal flaws in the data, in the treatment of the data or in the survey methodology are identified.

<u>Should data from the NMFS longline survey be used in either assessment?</u>

The AFSC longline survey was developed as an index of abundance for sablefish, but recently, the results have also been found useful as indices of abundance for rougheye rockfish, blackspotted rockfish, and for black halibut (aka Greenland turbot). Similar to the IPHC longline survey, the AFSC longline survey started in 1997, but it is conducted during odd years in the Eastern Bering Sea and during even years in the Aleutian Islands. In the Gulf of Alaska, the survey is conducted every year. There are few stations in the Eastern Bering Sea and in the Aleutian Islands, but they do cover the expected area of distribution of larger Pacific cod (Figure 5).

The agreement between the AFSC shelf trawl survey and the AFSC longline survey (Figure 6) is better than between the AFSC shelf trawl survey and the IPHC longline survey. The AFSC longline survey, similar to the IPHC longline survey, catch different sizes than the AFSC shelf trawl survey (Figure 7).

The agreement between the AFSC longline survey and the IPHC longline survey (Figure 8) is poor overall. The two apparently anomalous points in the IPHC longline survey in 1999 and 2005 may explain in part the discrepancy, but differences in the area surveyed, in the timing of the survey and slight differences in the size composition may also play a role (keeping in mind that further work may be needed to confirm that the index derived from the IPHC longline survey is appropriate).

Similar to the IPHC longline survey, there are no objective reasons to reject the AFSC longline survey as an index of stock size. The AFSC longline survey should also be thoroughly investigated and used in the assessment unless fatal flaws in the data, in the data treatment or in the survey methodology are identified.

It is only by analyzing additional data that confidence will increase in the model results. Given the widely different results that can be obtained with SS3 (Figure 4), and the volatility of some of those results, it is not be possible to model oneself out of the uncertainty. Only careful examination and inclusion of informative additional data will allow that.

The discussion above is based on examination of data from the Eastern Bering Sea surveys, but the conclusions and recommendations also hold for the Aleutian Islands data and assessment.

<u>How should the various data sets be weighted?</u>

Stock Synthesis is a very flexible stock assessment framework. Giving different weights to the various data sources, and depending on assumptions (e. g. fixed parameters), very different results in terms of absolute stock size, but also sometimes in terms of trends, can be obtained (Figure 4). This can also occur with other assessment frameworks, but because of SS3's flexibility, the problem is more severe.

Generally speaking, indices of abundance should be given more weight in the assessment than length composition. Age composition, particularly from surveys or other indices of abundance can be very informative if analyzed and used appropriately. Information in the length composition is at best indirect information on changes in stock size and it may be misleading if substantial changes in growth occur over time (Figure 11). In almost every stock where growth information is available by year, growth has been found to vary with trends over time, sometimes quite considerably. SS3 does allow for time varying growth, but without external information, it is unlikely to be able to estimate changes in growth correctly.

<u>What form (i.e., Stock Synthesis “pattern”) should be used for the selectivity functions?</u>

Selectivity is a very important parameter in any assessment framework. Changes in growth, natural mortality, or fishing mortality can all be aliased as changes in selectivity. Several of the model configurations examined during the review had very peculiar selectivity patterns. This was identified in the pre-review material and in the presentations by the assessment team. Those were probably not real and were likely due to sampling problems or aliasing other changes. My preference would be to NOT allow too much flexibility in selectivity changes over time, and to NOT allow strange patterns (e.g. figures 2.1.3 in the Eastern Bering Sea and 2A.11 and 2A.12 in the Aleutian Islands in the December 2015 SAFE report). If allowing these strange patterns is a condition of getting a good fit or convergence, this would be a sign that something else might be wrong. If allowed to change over time and age, the changes should be relatively smooth and not result in peculiar patterns.

During the meeting, Robin Cook noted that the ratio of catch at length in the longline commercial fishery to the survey catch at length is reasonably constant above a certain length. A possible interpretation is that domed selectivity estimated for the AFSC shelf survey could be an artifact. Data in the file "Long-term sizecomp comparison (trawl surveys, longline surveys, longline fishery).xlsx", the ratio of the longline commercial catch at length to the various population estimates at length from the surveys are plotted in Figure 9. The ratio of the longline commercial catch at length to the ASFC shelf survey for the Eastern Bering Sea is consistent with the observed size composition (Figure 10). The longline commercial fishery catches very few Pacific cod less that 40cm and the ratio increases progressively from nearly zero at 40 cm to around 5-6 at 70 cm and the ratio is indeed relatively constant from 70cm or so. The ratio seems to decrease above 100cm but this could easily be the result of low sample size. The ratio being relatively constant at 70cm and above suggests that selectivity does not decrease at those sizes in the AFSC shelf trawl survey, or that selectivity in the longline commercial fishery decreases at a similar rate. This is unlikely but not impossible. The link between selectivity in the AFSC shelf trawl survey and selectivity in the longline commercial fishery should be further investigated to guide modeling.

The longline surveys appear to have very low selectivity, lower than that of the commercial longline fishery (Figure 9), for size less than the 55 cm for the IPHC longline survey and less than 40 cm for the AFSC longline survey. Differences in selectivity between the two longline surveys may be due to the differences in the sizes they catch (Figure 7). The ratio of the commercial longline catch at length to the survey catch at length is near 1 for both surveys in the 60-70 cm range, but the ratios diverge thereafter. The IPHC longline survey appears to have higher selectivity for the larger size than the commercial longline fishery does, while the AFSC longline survey would have lower selectivity than the commercial longline fishery. The AFSC slope trawl survey shows a pattern similar to the AFSC longline survey. The reason(s) for the apparent differences in selectivity between the IPHC longline survey and the AFSC longline survey for lengths above 70cm should be further investigated.

This is but a quick examination of what the data are telling us in terms of selectivity. Modeling results would be expected to be consistent with those observations.

<u>Should the models be structured with respect to season?</u>

In both areas, there seem to be a strong seasonal pattern in the fishery. Therefore, where the data are sufficient, it would be appropriate to structure the assessment model by season. However, for the Aleutian Islands assessment, the data may not be sufficient to structure by season.

<u>Should the models be structured with respect to gear type?</u>

Bottom trawl and longline are the two main gear types in the fisheries. Their size selectivities are expected to be different, and the models should definitely be structured with respect to gear type where the data are sufficient to do so.

<u>How much time variability should be allowed, and in which parameters?</u>

Selectivity, catchability of the surveys, natural mortality, and growth could be allowed to vary over time when there is independent information supporting that changes are happening. A change in the ratio between total catch biomass and biomass estimate in the survey that could not be explained by changes in management could be an indication that the catchability in the survey has changed. Changes in mesh sizes in the trawl or hook size in the longline fishery could be an indication of a stepped change in selectivity. Changes in the predator field or extreme weather events could be indications of changes in natural mortality. Because most of these parameters are interlinked, great care should be taken in allowing them to vary. Only those parameters where there is external information suggesting that changes are occurring should be allowed to vary, probably one at a time to avoid incorrect interpretation. Because of the flexibility in SS3 and because most of these parameters are interlinked allowing them to change may give strange results, such as highly anomalous selectivity.

<u>What constraints, if any, should be placed on survey selectivity at older ages?</u>

Peculiar selectivity patterns have been identified as a problem in the presentations by the assessment team. Based on the information in Figure 9, the selectivity for the AFSC shelf trawl survey in the Eastern Bering Sea at ages corresponding to 70 cm and larger would be expected to be reasonably flat. For both areas, sharp peak and valleys, unless based on external information, should be smoothed. As indicated above, strange, irregular patterns should be constrained to be smoother.

<u>What constraints, if any, should be placed on survey catchability?</u>

In the mid 1980s survey catchability was estimated for cod and haddock on the Eastern Scotian Shelf and in the Gulf of St. Lawrence in Eastern Canada. Catchability for cod at the time was about 0.5 and for haddock it was close to 1. Vessels and gears have changed since and catchability estimates have also changed and in some areas they are now estimated to be greater than 1. Survey catchability is a scaling factor. In most assessment in the ICES area, survey catch per tow or catch per hour are used in the assessments and survey catchability is not an issue. It is, however, good practice to check every now and then if the assessment has the units more or less right by doing the areal expansion and comparing with the population estimates in the assessment.

Survey catchability smaller than 1 are relatively easy to rationalize e.g. by fish swimming faster than the net is towed, escaping above or below the net, or being more abundant in areas that are not surveyed. Survey catchability greater than 1 would happen if there is herding or if fish density in the surveyed area is expanded to areas where there are no fish, e.g. expanding flatfish density estimates from samples on smooth flatfish habitat to rough hard substrate that are not sampled and where flatfish are not present.

Catchability and natural mortality are interlinked. Considerable work was done in the Gulf of St. Lawrence in eastern Canada following the collapse of the groundfish stocks. Sinclair (20014) estimated that natural mortality had likely increased for the southern Gulf of St. Lawrence stock. Subsequent stock assessments (e.g. http://www.dfo-mpo.gc.ca/csas-sccs/Publications/ResDocs-DocRech/2007/RES2007_033_B.pdf and http://www.dfo-mpo.gc.ca/csas-sccs/Publications/ResDocs-DocRech/2007/RES2007_068_B.pdf) have used time varying natural mortality, but Canadian scientists warned that "Estimation of M can be confounded by changes in survey catchability and fishery catch reporting, and may be sensitive to assumptions and constraints applied in the ADAPT estimation procedure." (http://www.dfo-mpo.gc.ca/csas/Csas/status/2007/SAR-AS2007_002_E.pdf). Therefore, estimating catchability and natural mortality simultaneously would be challenging in the absence of external information.

External information indicative of changes in catchability could be changes in gears in the surveys or changes in predator abundance. Changes in catchability of pelagic species has been hypothesized to explain apparent increases of small pelagics in Eastern Canada after the collapse of groundfishes but this has been challenged. Catchability in longline surveys could occur if high prey abundance in the water decreases the attractiveness of baited hooks.

This being said, Pacific cod appears to be a relatively well behaved species as far as trawl surveys are concerned. Survey catchability estimates between 0.5 and 1.5 would not seem to be cause for concern. The assessment team, the PDT and the SSC are concerned that catchability less than 1 imply very large biomass estimates. As indicated above, I do not share that concern (within limits of course). Catchability of the trawl survey in the Aleutian Islands area would be expected to be more uncertain than in the Eastern Bering Sea area because bottom topography is likely rougher and more diverse in the Aleutian Islands area than in the Eastern Bering Sea area.

How should large gradients be dealt with in otherwise apparently converged models?

Stock Synthesis User Manual version 3.24s, page 27, states: "When using more population length bins than data bins, SS will run slower (more calculations to do), the calculated weights at age will be less aliased by the bin structure, and you may or may not get better fits to your data.

While exploring the performance of models with finer bin structure, a potentially pathological situation has been identified. When the bin structure is coarse (note that some applications have used 10 cm bin widths for the largest fish), it is possible for a selectivity slope parameter or a retention parameter to become so steep that all of the action occurs within the range of a single size bin. In this case, the model will lose the gradient of the logL with respect to that parameter and convergence will be hampered. A generic guidance to avoid this situation is not yet available."

I have no further advice on how to deal with large gradient than what is said in the Stock Synthesis User Manual.

Changes in growth

For the Eastern Bering Sea, weights at age in the survey (from the preliminary assessment data file) show trends over time that seem to be year-class specific. It could be worth investigating further changes in growth (Figure 11), particularly with respect to the implications for the assessment as growth changes may have an influence on fishing mortality and population estimates.

Recruitment index

For the Eastern Bering Sea, the population estimates in the AFSC shelf trawl survey seem to be reasonably consistent for the first 3 age groups or so with reasonably good year-class tracking (Figure 12). If the AFSC shelf trawl survey for the Eastern Bering Sea is indeed following year-classes reasonably well, it could provide at least 3 successive estimates of year-class size and this could be used to obtain reasonably reliable estimates of year-class sizes.

In my cursory comparison of the AFSC shelf trawl survey length frequencies with the age frequencies in the same survey, I got the impression that the smallest modal length group was sometimes aged as age 1 and in other cases as age zero. This should be verified.

Exploitation rate

The ratio of the commercial catch in tons to the survey biomass estimate in tons should be an indication of exploitation rate (relative if the catch and survey biomass are not in the same units). Figure 13, using data from run 15.6 for the Eastern Bering Sea shows the catch/survey ratio in biomass compared with the fishing mortality estimate for the same model. The results suggest that fishing mortality in model 15.6 could be overestimated in recent years. Unless the catchability of the survey has changed, the results below suggest that F has been lower than average since about 2007. The correlation between the catch/survey ratio and F is low (0.009).

Reliability of total catch estimates

For the Aleutian Islands assessment model 15.7, there is reasonably good agreement between fishing mortality estimates in the assessment and catch (Figure 14) except in the late 1980s and in 2010 when fishing mortality estimates suggests that mortality has been higher. It might be worth investigating if additional sources of mortality (e.g. increased M) occurred in those years. The correlation between F from the assessment (model 15.7) and the ratio of catch to survey biomass is higher (Figure 15) for the Aleutian Islands (0.47).

Stock structure

In the Aleutian Islands area, it is unlikely that there is a single stock in the traditional understanding of the concept. Instead, a number of local spawning would be expected with limited mixing during spawning. While these different spawning units may react similarly to changes in the environment and show similar trends in recruitment, they are unlikely to form a single homogeneous biological unit. It is likely impractical to do individual stock assessments for each of the individual units, and lumping all units into a single assessment with indices of abundance for only a few of them may increase the risk to less productive units. Simpler form of monitoring and management, in close cooperation with the industry and possibly NGOs, could be a better way of protecting the resources and managing the fisheries.

Conclusions and Recommendations in accordance with the ToRs

For the IPHC longline survey, the appropriateness of the data and how it was treated to calculate an index should be further verified between now and the assessment meeting later in the year. What data were used and how they were used should also be documented. The apparent anomalies in 1999 and 2005 warrant further investigations to try to identify what might cause them. If there are valid reasons to exclude those two points, it might be possible to reconcile the IPHC longline and AFSC shelf trawl survey time series taking into account that they sample different size groups. From what was discussed during the meeting and the documentation reviewed, there are no objective reasons to reject the IPHC longline survey as an index of stock size, assuming it has been correctly put together and calculated. The IPHC longline survey data should be thoroughly investigated. It should be used in the assessment unless fatal flaws in the data, in the treatment of the data or in the survey methodology are identified.

Similar to the IPHC longline survey, there are no objective reasons to reject the AFSC longline survey as an index of stock size. The AFSC longline survey should also be thoroughly investigated and used in the assessment unless fatal flaws in the data, in the data treatment or in the survey methodology are identified.

The discussion above is based on examination of data from the Eastern Bering Sea surveys, but the conclusions and recommendations also hold for the Aleutian Islands data and assessment.

With respect to weighting different data sets, indices of abundance should be given more weight in the assessment than length composition. Age composition, particularly from surveys or other indices of abundance can be very informative if analyzed and used appropriately. Information in the length composition is at best indirect information on changes in stock size and it may be misleading if substantial changes in growth occur over time (Figure 11).

Regarding the form of the selectivity function, my preference would be to NOT allow too much flexibility in selectivity changes over time and to NOT allow strange patterns (e.g. figures 2.1.3 in the Eastern Bering Sea, and 2A.11 and 2A.12 in the Aleutian Islands in the December 2015 SAFE report). If allowing these strange patterns is a condition of getting a good fit or convergence, this would be a sign that something else might be wrong. If allowed to change over time and age, the changes should be relatively smooth and not result in peculiar patterns. The ratio of catch at length in the longline commercial fishery to the survey catch at length being relatively constant at 70cm and above suggests that selectivity does not decrease at those sizes in the AFSC shelf trawl survey, or that selectivity in the longline commercial fishery decreases at a similar rate. This is unlikely but not impossible. The link between selectivity in the AFSC shelf trawl survey and selectivity in the longline commercial fishery should be further investigated to guide modeling. The reason(s) for the apparent differences in selectivity between the IPHC longline survey and the AFSC longline survey for lengths above 70cm should be further investigated.

Where the data are sufficient, it would be appropriate to structure the assessment model by season.

Bottom trawl and longline are the two main gear types in the fisheries. Their size selectivity are expected to be different and the models should definitely be structured with respect to gear type where the data are sufficient to do so.

Selectivity, catchability of the surveys, natural mortality, and growth could be allowed to vary over time when there is independent information supporting that changes is happening. Because most of these parameters are interlinked, great care should be taken in allowing them to vary. Only those parameters where there is external information suggesting that changes is occurring should be allowed to vary, probably one at a time to avoid incorrect interpretation.

Based on the information in Figure 9, the selectivity for the AFSC shelf trawl survey in the Eastern Bering Sea at ages corresponding to 70 cm and larger would be expected be reasonably flat. For both areas, sharp peaks and valleys, unless based on external information, should be smoothed. As indicated above, strange, irregular patterns should be constrained to be smoother.

I have no further advice on how to deal with large gradient than what is said in the Stock Synthesis User Manual.

It could be worth investigating further changes in growth (Figure 11), particularly with respect to the implications for the assessment as growth changes may have an influence on fishing mortality and population estimates.

If the AFSC shelf trawl survey for the Eastern Bering Sea is indeed following year-classes reasonably well, it could provide at least 3 successive estimates of year-class size and this could be used to obtain reasonably reliable estimates of year-class sizes.

In my cursory comparison of the AFSC shelf trawl survey length frequencies with the age frequencies in the same survey, I got the impression that the smallest modal length group was sometimes aged as age 1 and in other cases as age zero. This should be verified.

Figure 13, using data from run 15.6 for the Eastern Bering Sea shows the catch/survey ratio in biomass compared with the fishing mortality estimate for the same model. The results suggest that fishing mortality in model 15.6 could be overestimated in recent years.

For the Aleutian Islands assessment model 15.7, there is reasonably good agreement between fishing mortality estimates in the assessment and catch (Figure 14) except in the late 1980s and in 2010 when fishing mortality estimates suggests that mortality has been higher. It might be worth investigating if additional sources of mortality (e.g. increased M) occurred in those years.

In the Aleutian Islands area, it is unlikely that there is a single stock in the traditional understanding of the concept. Simpler form of monitoring and management, in close cooperation with the industry and possibly NGOs, could be a better way of protecting the resources and managing the fisheries.

One cannot model oneself out of lack of data, particularly for the Aleutian Islands assessment. Stock Synthesis has so much flexibility that, given sufficient time, a skilled user can probably get almost any stock trend from a dataset. Indices of abundance should be given more weight in the assessment than length composition. Age composition, particularly from the commercial fishery, but also from surveys or other indices of abundance can be very informative if analyzed appropriately. Information in the length composition is at best indirect information on changes in stock size. In almost every stock where growth information is available by year, growth has been found to vary with trends over time, sometimes quite considerably and this could very well be the case here for the Eastern Bering Sea (Figure 11). SS3 does allow for time varying growth, but without external information, it is unlikely to be able to estimate changes in growth correctly.

Analytical retrospective analyses are routinely done for both stocks. Historical retrospective, where the successive accepted assessment are also informative and should be done to indicate how consistent the assessments have been over time.

Simpler models, e.g. like Robin Cook's or surplus production models should be investigated. It is not necessary to go to Ensemble modeling, but looking at more than one modeling framework might be informative.