Minutes of the Joint Team Subcommittee on Pacific Cod Models

May 16, 2016

Beginning with the 2010 assessment cycle, the BSAI and GOA Groundfish Plan Teams ("Joint Teams") have met by WebEx teleconference in the spring of each year to provide initial review of proposals for models to be included in the respective year's preliminary assessments of the various Pacific cod stocks. From 2010-2013, the full Joint Teams participated in these meetings. However, beginning with the 2014 assessment cycle, the Joint Teams delegated this responsibility to a Joint Team Subcommittee (JTS).

This year's JTS meeting took place on May 6, 2016. The JTS consisted of BSAI Team co-chair Dana Hanselman, GOA Team co-chair Jim Ianelli, and GOA Team member Sandra Lowe. All members were present, as were Grant Thompson (BSAI Team co-chair and senior author of the BS and AI Pacific cod assessments) and Chad See (Freezer Longline Coalition). Grant was appointed as rapporteur. In a departure from previous years' meetings, this year's meeting did not include proposals for the assessment of Pacific cod in the GOA, per request of the new senior author of that assessment, Steve Barbeaux.

The JTS noted that the BSAI Team currently includes only two members who conduct age-structured stock assessments, one of whom is precluded from being a JTS member by virtue of the fact that he is also the senior author of the assessments for the BS and AI Pacific cod stocks.

The JTS recommended that the SSC appoint additional members to the BSAI Team with expertise in conducting age-structured assessments.

One week prior to the meeting, JTS members were directed to the website for this year's CIE review of the BS and AI assessments (http://tinyurl.com/Pcod-cie-2016), which contains every file vetted during the review process as well as the final reports from the three reviewers. JTS members were also provided with a rough draft of Grant's summary of the CIE review (the final draft of which is included as Appendix 1 to these minutes). A total of 135 recommendations were provided by the CIE reviewers.

Grant began the meeting by giving a presentation on last year's assessments (both preliminary and final) in the two regions and a presentation on the CIE review, after which the JTS moved into deliberations on the various recommendations that had been developed since completion of last year's assessments.

In November/December of last year, the Team/SSC made fewer recommendations than usual regarding this year's assessments, in anticipation of the CIE review. Specifically, the BSAI Team made no recommendations, and the SSC made only six (the first of which is inferred, based on standing practice):

1. Standing request (both areas): *Include current base model*.
2. December 2015 minutes (Bering Sea only): "The SSC was encouraged by the author's explanation that dome-shaped selectivity may, in part, be explained by the possibility that some of older fish may be residing in the northern Bering Sea (NBS) at the time of the survey. This is supported by the size composition of the fish in the 2010 NBS trawl survey, which suggested that up to 40% of the fish in some larger size classes reside in this area, although the overall proportion in the NBS was small. The SSC encourages the author to further examine Pacific cod catches from trawl surveys conducted triennially by the National Marine Fisheries Service (NMFS) (1976-1991) and by the Alaska Department of Fish & Game (1996 to the present) to monitor the distribution and abundance of red king crab and demersal fish (see: Hamazaki, T., Fair, L., Watson, L., Brennan, E., 2005. Analyses of Bering Sea bottom-trawl surveys in Norton Sound: absence of regime shift effect on epifauna and demersal fish. ICES Journal of Marine Science 62, 1597-1602). While the 2010 bottom trawl survey in the NBS found relatively few

Pacific cod (3% of total biomass), it is possible that the proportion of Pacific cod that are outside the standard survey area was higher in other years. A second possibility is that older Pacific cod migrate to nearshore areas to feed in the summer, making them unavailable to the survey." Summary: *Examine NMFS and ADFG survey data from the northern BS and Norton Sound*.

3. December 2015 minutes (Bering Sea only): "The SSC noted that the iteratively tuned, time-varying parameters in the model have not been updated since 2009. The author confirmed that the currently assumed standard deviations of two dev vectors (log of age-0 recruitment and a parameter corresponding to the ascending part of the selectivity curve) may no longer match the standard deviations of these vectors, which could contribute to retrospective bias. The SSC looks forward to a new paper on this issue that the author is preparing." Summary: *Circulate manuscript on estimating standard deviations of time-varying parameters*.
4. December 2015 minutes (Bering Sea only): "While the model selection criteria proposed by the author are reasonable, we note that these criteria do not take into account the model fit itself. Model fit and retrospective performance should be more strongly considered in the selection of a final model for specifications." Summary: *Weight model fit and retrospective performance more heavily in selection criteria*.
5. December 2015 minutes (Bering Sea only): "Although the SSC has repeatedly stressed the need to incrementally evaluate model changes, the SSC did not intend this to imply an automatic preference for the status quo model (as implied by the authors criterion #1) if alternatives with better performance are available." Summary: *Evaluate model changes incrementally; do not automatically prefer base model*.

The JTS used the above list and Table 2 from Appendix 1 to structure its discussion and summarize its recommendations. The purpose of the recommendations was to winnow the lists of proposals into smaller sets of models and non-model analyses to be included in this year's preliminary assessments, with the understanding that the assessment authors can bring forward additional models and non-model analyses at any time.

During the discussion, the JTS developed 5 of its own recommendations:

1. Use empirical weight at age (Bering Sea only). This is an option in Stock Synthesis (SS), where a vector of weights at age is used instead of combined weight-at-length and length-at-age relationships. It has proven helpful in several assessments of groundfish on the west coast.
2. Include IPHC longline survey, with "extra SD" (both areas). Several CIE comments suggested that the IPHC longline survey data appear to be usable in both areas, although the reviewers suggested that further investigation of these data and possible issues regarding data weighting would be appropriate (comments 1a.01-1a.10), with some suggestions for removal of possible outliers (1a.11-1a.13). The "extra SD" is a feature in SS that allows the observation error standard deviations associated with a survey index time series to be estimated internally by adding a constant to the design-based standard deviations.
3. Include NMFS longline survey, with "extra SD" (both areas). Similar to the IPHC longline survey, several CIE comments suggested that the NMFS longline survey data also appear to be usable in both areas, although the reviewers again suggested that further investigation of these data and possible issues regarding data weighting would be appropriate (comments 1b.01-1b.08), with one suggestion for evaluation of a possible step-change since 2010 (1b.09).
4. Include IPHC and NMFS longline surveys, with "extra SD" for both (both areas). See previous two recommendations.
5. Use reasonably time-varying, double normal selectivity (Bering Sea only). CIE comments 2e.01 and 2e.09 suggested that some amount of time-variability in fishery selectivity is appropriate, CIE comment 2e.12 cautioned against allowing "too much" time-variability in selectivity, and CIE comment 2b.07 suggested use of the double normal selectivity function.

The discussion took place in three phases, or "passes." On the first pass, the JTS considered each item in the above list and every comment summary in Table 2 from Appendix 1, and rated it "yes" or "no" (meaning "do" or "do not" make it a priority to be considered before the next CIE review, anticipated for early 2021). On the second pass, the JTS rated the priority of each item receiving a "yes" on the first pass as follows: high = to be completed during this year's assessment, med = to be completed during the 2017 or 2018 assessments, and low = to be completed during the 2019 or 2020 assessments. On the third pass, the JTS assigned each high priority item to a model (maximum of 6 in each area, including the current base model). The results of this exercise are shown in Table 1. Note that the terms "comment" and "proposal" are used interchangeably here. *Note also that the model numbers (1-6) shown in Table 1 are only placeholders; the actual model numbers will be assigned during this summer's analysis.* Table 2 is the same as Table 2 from Appendix 1, except that comments identified as priorities by the JTS are highlighted, using the following color codes:

- Green means that the comment was ranked "high" in both areas.
- Yellow means that the comment was ranked "med" in both areas.
- Red means that the comment was ranked "low" in both areas.
- Grey means that the comment was ranked differently in the two areas (including cases where the comment was ranked in one area and unranked (i.e., first pass = "no") in the other).

The JTS anticipates that any comments currently ranked as "med" or "low" priority may be re-evaluated in the future.

For the BS, the subcommittee recommended that the following models be developed for this year's preliminary assessment:

- Model 1: BS Model 11.5, the final model from 2015 (same as the final models from 2011-2014)
- Model 2: Like BS Model 15.6, but simplified as follows:
    1. Weight abundance indices more heavily than sizecomps.
    2. Use the simplest selectivity form that gives a reasonable fit.
    3. Do not allow survey selectivity to vary with time.
    4. Do not allow survey catchability to vary with time.
    5. Force trawl survey selectivity to be asymptotic.
    6. Do not allow strange selectivity patterns.
    7. Use empirical weight at age.
- Model 3: Like BS Model 15.6, but including the IPHC longline survey data and other features, specifically:
    1. Do not allow strange selectivity patterns.
    2. Estimate catchability of new surveys internally with non-restrictive priors.
    3. Include additional data sets to increase confidence in model results.
    4. Include IPHC longline survey, with "extra SD."
- Model 4: Like Model 3 above, but including the NMFS longline survey instead of the IPHC longline survey.
- Model 5: Like Models 3 and 4 above, but including both the IPHC and NMFS longline survey data and two features not included in either Model 3 or 4, specifically:
    1. Start including fishery agecomp data.
    2. Use empirical weight at age.
- Model 6: Like Model 5 above, but including two features not included in Model 5, specifically:
    1. Use either Francis or harmonic mean weighting.
    2. Explore age-specific M (e.g., using Lorenzen function).

For the EBS, the JTS recommended that the following non-model analysis be conducted for this year's preliminary assessment:

- Non-model analysis 1: Verify that the trawl survey data sometimes include age 0 fish.

For the AI, the JTS recommended that the following models be developed for this year's preliminary assessment:

- Model 1: AI Model 13.4, the final model from 2015 (Tier 5 random effects model)
- Model 2: Like AI Model 15.7, but simplified as follows:
    1. Weight abundance indices more heavily than sizecomps.
    2. Use the simplest selectivity form that gives a reasonable fit.
    3. Do not allow survey selectivity to vary with time.
    4. Do not allow survey catchability to vary with time.
    5. Do not allow strange selectivity patterns.
    6. Estimate trawl survey catchability internally with a fairly non-informative prior.
- Model 3: Like AI Model 15.7, but including the IPHC longline survey data and other features, specifically:
    1. Do now allow strange selectivity patterns.
    2. Estimate trawl survey catchability internally with a fairly non-informative prior.
    3. Estimate catchability of new surveys internally with non-restrictive priors.
    4. Include additional data sets to increase confidence in model results.
    5. Include IPHC longline survey, with "extra SD."
- Model 4: Like Model 3 above, but including the NMFS longline survey instead of the IPHC longline survey.
- Model 5: Like Models 3 and 4 above, but including both the IPHC and NMFS longline survey data.
- Model 6: Like AI Model 15.7, except:
    1. Use the post-1994 AI time series (instead of the post-1986 time series).
    2. Do not allow strange selectivity patterns.
    3. Estimate trawl survey catchability internally with a fairly non-informative prior.

With respect to AI Model 2, the JTS noted that CIE comments 2i.16-2i.19 also suggest exploring "simple" models for the AI stock, although these CIE comments all pertain to use of models developed outside of SS, whereas the JTS recommended instead that a "simple" model be developed within the SS framework.

For the AI, the JTS did not recommend any non-model analyses for this year's preliminary assessment.

The JTS recognized that some of the terms used in the above recommendations are subjective and that, in making these recommendations, the assessment author will need to determine:

1. How to measure the weight assigned to abundance indices and sizecomp data in the same units.
2. What constitutes a "reasonable fit."
3. What constitutes a "strange" selectivity pattern.
4. What constitutes a "fairly non-informative" prior.

Table 1. Proposals ranked as either high, medium, or low by the JTS, with high priority proposals assigned to September models for both areas.

Note: September model numbers are temporary placeholders; actual numbers will be established during analysis.
Abbreviations: AI = Aleutian Islands, BS = Bering Sea, D = author's discretion, JTS = Joint Team Subcommittee, n/a = not applicable, NMA = non-model analysis, No. = proposal number, Pri. = priority, SPM = starting point model, SSC = Scientific and Statistical Committee

		Bering Sea								Aleutian Islands							
No.	Brief description of proposal	Pri.	SPM	1	2	3	4	5	6	NMA	Pri.	SPM	1	2	3	4	5	6	NMA
SSC1	Include current base model	high	11.5	x							high	13.4	x						
1.05	Use the post-1994 AI trawl survey time series	n/a									high	15.7						x	
2a.07	Use either Francis or harmonic mean weighting	high	15.6						x		n/a								
2a.08	Weight abundance indices more heavily than sizecomps	high	15.6		x						high	15.7		x					
2b.08	Use the simplest selectivity form that gives a reasonable fit	high	15.6		x						high	15.7		x					
2e.11	Do not allow survey selectivity ... to vary with time	high	15.6		x						high	15.7		x					
2e.18	Do not allow survey ... catchability to vary with time	high	15.6		x						high	15.7		x					
2f.03	Force trawl survey selectivity to be asymptotic	high	15.6		x						n/a								
2f.06	Do not allow … "strange" selectivity patterns	high	15.6		x	x	x	x	x		high	15.7		x	x	x	x	x	
2g.03	Estimate catchability internally with a "fairly non-informative" prior	n/a									high	15.7		x	x	x	x	x	
2g.04	Estimate catchability of new surveys internally with non-restrictive priors	high	15.6			x	x	x	x		high	15.7			x	x	x		
2i.06	Explore age-specific M (e.g., using Lorenzen function)	high	15.6						x		n/a								
2i.13	Verify that the trawl survey data sometimes include age 0 fish	high	15.6							x	n/a								
2i.38	Include additional data sets to increase confidence in model results	high	15.6			x	x	x	x		high	15.7			x	x	x		
2i.39	Start including fishery agecomp data	high	15.6					x	x		n/a								
JTS1	Use empirical weight at age	high	15.6		x			x	x		n/a								
JTS2	Include IPHC longline survey, with "extra SD"	high	15.6			x					high	15.7			x				
JTS3	Include NMFS longline survey, with "extra SD"	high	15.6				x				high	15.7				x			
JTS4	Include IPHC and NMFS longline surveys, with "extra SD" for both	high	15.6					x	x		high	15.7					x		
SSC2	Examine NMFS and ADFG survey data from the northern BS and Norton Sound	med									n/a								
2a.07	Use either Francis or harmonic mean weighting	n/a									med								
2b.03	Investigate alternatives to double-normal selectivity	n/a									med								
2e.06	Allow time variability only where supported by external data	med									med								
2i.17	Investigate whether a simpler (than SS) model would be useful	n/a									med								
JTS5	Use reasonably time-varying, double normal selectivity	med									n/a								
2c.01	Use annually varying selectivity if it fits as well as season/gear structure	n/a									low								
2e.21	Consider time-varying growth if supported by data	low									low								
2g.03	Estimate catchability internally with a "fairly non-informative" prior	low									n/a								
2i.04	Do not include more model features than can be supported by the data	low									n/a								
2i.39	Start including fishery agecomp data	n/a									low								

Table 2. Summary of reviewer comments by ToR and area (page 1 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		1. Data currently used in the assessment models	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
1.01	AI	Explain yearly variation in AI catches using objective criteria	1	0	0	1
1.02	AI	Improve documentation for the AI trawl survey time series	0	1	0	1
1.03	AI	Reduce variability in catch or acknowledge that survey does not reflect biomass	1	0	0	1
1.04	AI	Use the post-1986 AI trawl survey time series	1	0	0	1
1.05	AI	Use the post-1994 AI trawl survey time series	0	1	0	1
1.06	both	Ask the data collectors to rank their data in order of potential bias	0	1	0	1
1.07	both	For all data sets, document how they meet specified criteria	0	1	0	1
1.08	both	Give primary responsibility for data documentation to the collectors	0	1	0	1
1.09	both	Include all data sets that meet specified criteria	0	1	0	1
1.10	both	Investigate all data more thoroughly	1	0	0	1
		1a. Potential use of IPHC longline survey data	**1**	**1**	**1**	**3**
		Examine IPHC survey in both areas more closely; use if no red flags	*1*	*1*	*1*	*3*
1a.01	BS	Examine IPHC survey data more thoroughly before using	0	1	1	2
1a.02	BS	Include IPHC survey as a means of stabilizing the BS assessment	0	1	0	1
1a.03	BS	Include the IPHC longline survey in order to explain cryptic biomass	0	1	0	1
1a.04	BS	Include IPHC survey to provide information on larger fish	0	1	1	2
1a.05	BS	Use IPHC longline and slope trawl surveys if weighted appropriately	1	0	0	1
1a.06	BS	Include the IPHC longline survey and slope trawl survey	1	0	0	1
1a.07	AI	Examine IPHC survey data more thoroughly before using	0	1	1	2
1a.08	both	Examine IPHC survey data more thoroughly before using	0	1	0	1
1a.09	both	If the IPHC longline survey is shown to be unbiased, use it	0	1	0	1
1a.10	both	If IPHC survey passes further investigation, use it	0	0	1	1
		Exclude certain years from IPHC time series in the BS if appropriate	*0*	*1*	*1*	*2*
1a.11	BS	Consider eliminating suspect data from the IPHC survey in the BS	0	1	0	1
1a.12	BS	Exclude 1999 and 2005 IPHC indices in the BS if appropriate	0	0	1	1
1a.13	BS	Investigate anomalous 1999 and 2005 IPHC indices in the BS	0	1	0	1
		Other	*0*	*0*	*1*	*1*
1a.14	BS	Investigate selectivity differences between the two longline surveys	0	0	1	1
		1b. Potential use of NMFS longline survey data	**1**	**1**	**1**	**3**
		Examine NMFS longline in both areas more closely; use if no red flags	*1*	*1*	*1*	*3*
1b.01	BS	Examine NMFS longline survey data more thoroughly before using	0	1	0	1
1b.02	BS	Include the NMFS longline survey in order to explain cryptic biomass	0	1	0	1
1b.03	BS	Include the NMFS longline survey to provide information on larger fish	0	0	1	1
1b.04	BS	Use NMFS longline and slope trawl surveys if weighted appropriately	1	0	0	1
1b.05	BS	Include the NMFS longline survey and slope trawl survey	1	0	0	1
1b.06	AI	Examine NMFS longline survey data more thoroughly before using	0	1	0	1
1b.07	both	If the NMFS longline survey is shown to be unbiased, use it	0	1	0	1
1b.08	both	If NMFS longline survey passes further investigation, use it	0	0	1	1
		Other	*0*	*1*	*0*	*1*
1b.09	BS	Evaluate possible bias of NMFS longline survey since 2010	0	1	0	1

Table 2. Summary of reviewer comments by ToR and area (page 2 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2a. How should the various data sets be weighted?	**1**	**1**	**1**	**3**
		Weighting of data in general	*1*	*1*	*1*	*3*
2a.01	BS	Test weighting sensitivity by assuming agecomp is lognormal	1	0	0	1
2a.02	both	Do not use arithmetic mean weighting	0	1	0	1
2a.03	both	Downweight conflicting indices objectively	0	1	0	1
2a.04	both	Downweight sizecomps, as they are only indirect measures of abundance	0	0	1	1
2a.05	both	Downweight sizecomps, as they can mislead if growth changes over time	0	0	1	1
2a.06	both	Set multinomial input N equal to number of sampled trips	0	1	0	1
2a.07	both	Use either Francis or harmonic mean weighting	0	1	0	1
2a.08	both	Weight abundance indices more heavily than sizecomps	0	0	1	1
2a.09	both	Weight agecomps heavily	0	0	1	1
2a.10	both	Weight all indices internally (SS "extra SD")	0	1	0	1
		Weighting of IPHC and NMFS longline survey data in particular	*1*	*0*	*0*	*1*
2a.11	BS	Address issue of how to weight IPHC and NMFS longline surveys	1	0	0	1
2a.12	BS	Investigate whether including both longline surveys overweights large fish	1	0	0	1
2a.13	BS	Weight IPHC and NMFS longline surveys internally (SS "extra SD")	1	0	0	1
		2b. What form (SS "pattern") should be used for the selectivity functions?	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
2b.01	BS	Fix fishery selectivity above age 8	1	1	0	2
2b.02	BS	Use random-walk,/time-varying selectivity	1	0	0	1
2b.03	AI	Investigate alternatives to double-normal selectivity	0	1	0	1
2b.04	AI	Investigate sensitivity of results when using parametric selectivity	1	0	0	1
2b.05	AI	Use a parametric selectivity function	1	0	0	1
2b.06	both	Test random-walk/time-varying vs. parametric/partitioned selectivity	0	1	0	1
2b.07	both	Use double-normal selectivity unless there is good reason not to	0	1	0	1
2b.08	both	Use the simplest selectivity form that gives a reasonable fit	0	1	0	1
		2c. Should the models be structured with respect to season?	**0**	**1**	**1**	**2**
		Various	*0*	*1*	*1*	*2*
2c.01	BS	Use annually varying selectivity if it fits as well as season/gear structure	0	1	0	1
2c.02	AI	Consider whether data are sufficient to support seasonal structure	0	0	1	1
2c.03	both	Structure models by season if data are sufficient	0	0	1	1
		2d. Should the models be structured with respect to gear?	**0**	**0**	**1**	**1**
		General (some reviewers merged ToR 2d comments into ToR 2b and 2c)	*0*	*0*	*1*	*1*
2d.01	both	Structure models by gear if data are sufficient	0	0	1	1

Table 2. Summary of reviewer comments by ToR and area (page 3 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2e. How much time variability should be allowed, and in which parameters?	**1**	**1**	**1**	**3**
		General	*1*	*1*	*1*	*3*
2e.01	BS	Limit time variability to recruitment and selectivity	0	1	0	1
2e.02	BS	Remove time variability from parameters exhibiting little variation	0	1	0	1
2e.03	BS	Use time blocks only if supported by prior data examination or external data	0	1	0	1
2e.04	AI	Consider simplifying model to improve retrospective behavior	0	1	0	1
2e.05	AI	Do not use random effects for a survey with less than annual data	1	0	0	1
2e.06	both	Allow time variability only where supported by external data	0	0	1	1
2e.07	both	Consider how time-varying parameters affect reference point estimation	0	1	0	1
		Time-varying selectivity	*1*	*1*	*1*	*3*
2e.08	BS	Changes in mixed-gear selectivity should mirror ratios of catch by gear	0	1	0	1
2e.09	BS	Model time variability in selectivity as a random walk	1	0	0	1
2e.10	both	Account for trends in survey selectivity ... outside the model	0	1	0	1
2e.11	both	Do not allow survey selectivity ... to vary with time	0	1	0	1
2e.12	both	Do not allow "too much" time-variability in selectivity…	0	0	1	1
2e.13	both	Allow fishery selectivity to vary fairly freely if model is not gear-structured	0	1	0	1
		Time-varying catchability	*1*	*1*	*0*	*2*
2e.14	BS	Consider allowing time variability in catchability carefully	1	0	0	1
2e.15	BS	Investigate residual "runs" in models with time-varying catchability	0	1	0	1
2e.16	BS	Model time variability in catchability as a random walk	1	0	0	1
2e.17	both	Account for trends in survey ... catchability outside the model	0	1	0	1
2e.18	both	Do not allow survey ... catchability to vary with time	0	1	0	1
		Time-varying growth	*0*	*1*	*1*	*2*
2e.19	BS	Consider implications of time-varying growth	0	0	1	1
2e.20	BS	Use external data to help SS estimate time-varying growth	0	0	1	1
2e.21	both	Consider time-varying growth if supported by data	0	1	0	1
		2f. What constraints, if any, should be placed on old-age survey selectivity?	**0**	**1**	**1**	**2**
		Various	*0*	*1*	*1*	*2*
2f.01	BS	Consider relationship between dome-shaped survey selectivity and M	0	1	0	1
2f.02	BS	Estimate survey selectivity at older ages freely, with sensitivity analysis	0	1	0	1
2f.03	BS	Force trawl survey selectivity to be asymptotic	0	0	1	1
2f.04	BS	Match trawl survey selectivity to longline fishery selectivity at large sizes	0	0	1	1
2f.05	both	Accept dome-shaped survey selectivity only if it occurs in many models	0	1	0	1
2f.06	both	Do not allow … "strange" selectivity patterns	0	0	1	1
		2g. What constraints, if any, should be placed on survey catchability?	**1**	**1**	**1**	**3**
		Use a prior distribution for catchability	*0*	*1*	*1*	*2*
2g.01	BS	Develop a prior for survey catchability "that can be generally agreed"	0	1	0	1
2g.02	AI	Expect AI survey catchability to be estimated less precisely than in the BS	0	0	1	1
2g.03	both	Estimate catchability internally with a "fairly non-informative" prior	0	1	0	1
2g.04	both	Estimate catchability of new surveys internally with non-restrictive priors	0	1	0	1
		Adjust bounds on catchability	*0*	*1*	*1*	*2*
2g.05	both	Estimate catchability internally with "agreed bounds" (e.g, 0.5-2 ´ estimate)	0	1	0	1
2g.06	both	Estimate catchability internally with bounds (0.5,1.5)	0	0	1	1
		Other	*1*	*1*	*1*	*3*
2g.07	BS	Estimate catchability internally	1	0	0	1
2g.08	both	Examine, rank, and document all potential sources of error in catchability	0	1	0	1
2g.09	both	Occasionally compare model's estimated survey biomass to area-swept value	0	0	1	1
		2h. How should big gradients be dealt with in apparently converged models?	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
2h.01	BS	Investigate large gradient problem by using alternative minimizers	1	0	0	1
2h.02	both	Do not worry about large gradients if the Hessian matrix can be inverted	0	1	0	1

Table 2. Summary of reviewer comments by ToR and area (page 4 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2i. Other comments (reviewers' choice)	**1**	**1**	**1**	**3**
		General modeling philosophy	*1*	*1*	*1*	*3*
2i.01	AI	Conduct much more model exploration	1	0	0	1
2i.02	AI	Do not use SS unless further model exploration indicates otherwise	1	0	0	1
2i.03	both	Do not accept models with significant retrospective bias	0	1	0	1
2i.04	both	Do not include more model features than can be supported by the data	0	0	1	1
		Natural mortality rate	*1*	*1*	*1*	*3*
2i.05	BS	Do not estimate both M and catchability internally if catchability varies with time	1	0	0	1
2i.06	BS	Explore age-specific M (e.g., using Lorenzen function)	0	1	0	1
2i.07	BS	Fix M as an age-dependent vector	1	0	0	1
2i.08	BS	Fix M as an age-dependent vector, using Lorenzen relationship	1	0	0	1
2i.09	BS	Use "Piner plots" to identify data sources that are in conflict with estimated M	0	1	0	1
2i.10	AI	Investigate whether M was higher than usual in the late 1980s and in 2010	0	0	1	1
2i.11	both	Do not estimate both M and catchability internally	0	0	1	1
		Investigation of suspect results or model features	*1*	*0*	*1*	*2*
2i.12	BS	Investigate whether fishing mortality for recent years is overestimated	0	0	1	1
2i.13	BS	Verify that the trawl survey data sometimes include age 0 fish	0	0	1	1
2i.14	AI	Investigate significance of very large changes in annual catch biomass	1	0	0	1
2i.15	both	Investigate whether distribution of length at age in SS implies incremental growth	1	0	0	1
		Alternative models ("simple")	*1*	*0*	*1*	*2*
2i.16	AI	Include Robin Cook's "simple" model	1	0	0	1
2i.17	AI	Investigate whether a simpler (than SS) model would be useful	1	0	0	1
2i.18	AI	Use a simpler form of monitoring and management, involving industry and NGOs	0	0	1	1
2i.19	both	Investigate whether a simpler (than SS) model would be useful	0	0	1	1
		Retrospective diagnostics	*0*	*1*	*1*	*2*
2i.20	both	Include "historical" retrospectives (i.e., from previous assessments)	0	0	1	1
2i.21	both	Include "Ianelli squid plots" for time-varying parameters	0	1	0	1
		Features to add in SS and use once added	*1*	*1*	*0*	*2*
2i.22	BS	Explore length-based random walk selectivity if it becomes an option in SS	0	1	0	1
2i.23	both	Include calculation of DIC in SS	1	0	0	1
2i.24	both	Modify SS so that F is explicitly modeled as an age/size effect ´ a year effect	1	0	0	1
2i.25	both	Rewrite SS in ADMB RE	1	1	0	2
2i.26	both	Treat catch data in SS as observations as opposed to parameters	1	0	0	1
2i.27	both	Use Dirichlet multinomial likelihood if it becomes an option in SS	0	1	0	1
		Sensitivity testing	*0*	*1*	*0*	*1*
2i.28	BS	Highlight sensitivity of model results to assumed value of "steepness."	0	1	0	1
2i.29	both	Consider possible effects of 1989 and 1999 regime shifts	0	1	0	1
2i.30	both	Use alternative plausible historical catch scenarios to test sensitivity	0	1	0	1
		Alternative models (fully age- or length-structured)	*1*	*0*	*0*	*1*
2i.31	both	Consider alternative models with different assumptions about errors in the data	1	0	0	1
2i.32	both	Consider using a truly Bayesian approach	1	0	0	1
2i.33	both	Consider using SAM (Nielsen and Berg, 2014)	1	0	0	1
2i.34	both	Use a model that includes a length-based growth projection matrix	1	0	0	1
		Prior distributions and penalty functions	*1*	*0*	*0*	1
2i.35	both	Distinguish between Bayesian priors and penalty functions	1	0	0	1
2i.36	both	Give much more thought to choice of penalty functions, especially bounds	1	0	0	1
		Use of additional data (other than longline surveys)	*0*	*0*	*1*	*1*
2i.37	BS	Use trawl survey estimates at ages 1-3 as an index of recruitment	0	0	1	1
2i.38	both	Include additional data sets to increase confidence in model results	0	0	1	1
2i.39	both	Start including fishery agecomp data	0	0	1	1

Appendix 1

Summary of the 2016 CIE review of the stock assessments for Pacific cod in the eastern Bering Sea and Aleutian Islands

Grant G. Thompson

Resource Ecology and Fisheries Management Division
Alaska Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
7600 Sand Point Way NE., Seattle, WA 98115-6349

May 16, 2016

The stock assessments for Pacific cod in the eastern Bering Sea (BS) and Aleutian Islands (AI) were reviewed by representatives of the Center of Independent Experts (CIE) during a meeting held at the Alaska Fisheries Science Center in Seattle, Washington, during the dates February 16-19, 2016. The CIE reviewers were Robin Cook, Neil Klaer, and Jean-Jacques Maguire. The Terms of Reference (ToR) for the review are included here as Attachment 1, and the agenda, as adopted at the meeting, is included as Attachment 2. All documents associated with the review are available at the following website: http://tinyurl.com/Pcod-cie-2016.

Many models were evaluated by the reviewers. For the BS assessment, the reviewers examined Models 11.5 and 14.2 from the final 2015 BS assessment, Model 15.6 from the preliminary 2015 BS assessment (but updated so as to include the same data used in Model 14.2), and 17 new models (see the link labeled "List of Stock Synthesis models (Bering Sea)" on the website). For the AI assessment, the reviewers examined Model 15.7 from the final 2015 AI assessment and 10 new models (see the link labeled "List of Stock Synthesis models Aleutian Islands" on the website). I produced all of the new BS models for the review, and Steve Barbeaux graciously volunteered to produce all of the new AI models.

The reviewers' reports were received on April 18. Attachment 3 shows the main text of each of the three reports, excluding boilerplate language, with line numbers starting over at 1 for each reviewer (to see things like figures and footnotes, the full reports can be accessed on the review website). Highlighting shows places where I interpreted the text as constituting a recommendation (the alternating yellow-blue-yellow pattern is used just to help distinguish between recommendations located adjacent to one another in the text; note also that the terms "comment" and "recommendation" are used interchangeably here).

Table 1 lists, among other things, the line number(s) on which each comment contained highlighted in Attachment 3 begins. Sometimes a reviewer made the same comment more than once. I have therefore included the columns labeled "1st," "2nd," and "3rd," so that the locations of identical comments can be tracked (note that comment #129, by reviewer JM, actually appears in his report four times, with the fourth instance beginning on line 334). By "identical comments," I mean comments in the main text that are identical after correcting minor typos, reconciling minor punctuation differences, and (rarely) deleting superfluous text within the comment. The column labeled "Comment summary" lists my best attempt at summarizing the comments concisely. Note that the same comment summary sometimes applies to more than one row in the table (i.e., some comments, while not quite identical, are essentially equivalent). A total of 156 unique comments are identified in Table 1, but some of them are sufficiently similar that they map into only 127 comment summaries.

Table 2 lists the comment summaries from Table 1 in order of ToR and area (BS, AI, or both), with a new numbering system that links each comment summary to its respective ToR. In eight cases, the comment summaries from Table 1 were split into two parts for inclusion in Table 2, bringing the total number of comment summaries in Table 2 to 135. In five of these cases, the split was made in order to address the IPHC and NMFS longline surveys separately:

- The summary for unique comment #29 in Table 1 was split into summaries 1a.06 and 1b.05 in Table 2.
- The summary for unique comment #45 in Table 1 was split into summaries 1a.05 and 1b.04 in Table 2.
- The summary for unique comment #57 in Table 1 was split into summaries 1a.09 and 1b.07 in Table 2.
- The summary for unique comment #102 in Table 1 was split into summaries 1a.03 and 1b.02 in Table 2.
- The summary for unique comment #136 in Table 1 was split into summaries 1a.04 and 1b.03 in Table 2.

In two other cases, the split was made in order to address survey selectivity and survey catchability separately:

- The summary for unique comment #106 in Table 1 was split into summaries 2e.11 and 2e.18 in Table 2.
- The summary for unique comment #107 in Table 1 was split into summaries 2e.10 and 2e.17 in Table 2.

The final split was made in order to address selectivity time-variability and selectivity shape separately:

- The summary for unique comments #124 and #147 in Table 1 was split into summaries 2e.12 and 2f.06 in Table 2.

In Table 2, the columns labeled with the reviewers' initials show which reviewers made comments corresponding to the comment summary (multiple comments from the same reviewer corresponding to the same comment summary get the same score (1) as single comments). Table 2 also includes the (sometimes slightly abbreviated) text of the relevant ToR. Within each ToR, individual comment summaries are grouped under subheadings corresponding to ToR sub-themes. Sometimes the comments for a given ToR were so disparate that it was hard to find any common sub-themes, in which case there is just a single sub-theme labeled "Various." Headings (i.e., the text of the ToR) are shown in bold font, subheadings (sub-themes) are shown in italic font, and the comment summaries are shown in ordinary font. Where multiple sub-themes exist for a given ToR, they are listed in descending order of the number of reviewers that commented on them, except that whenever there was a sub-theme called "Other," I placed it last.

Table 1. Line numbers on which each unique comment begins, with comment summaries (page 1 of 3).

No.	Rev.	1st	2nd	3rd	Area	Comment summary
1	RC	5			BS	Use random-walk,/time-varying selectivity
2	RC	8			BS	Model time variability in selectivity as a random walk
3	RC	10	115		BS	Fix fishery selectivity above age 8
4	RC	14			BS	Fix M as an age-dependent vector
5	RC	15	129		BS	Model time variability in catchability as a random walk
6	RC	17	139		BS	Estimate catchability internally
7	RC	25	151		BS	Address issue of how to weight IPHC and NMFS longline surveys
8	RC	26	154		BS	Weight IPHC and NMFS longline surveys internally (SS "extra SD")
9	RC	30			BS	Test weighting sensitivity by assuming agecomp is lognormal
10	RC	38			AI	Use a parametric selectivity function
11	RC	39			AI	Investigate sensitivity of results when using parametric selectivity
12	RC	41			AI	Reduce variability in catch or acknowledge that survey does not reflect biomass
13	RC	46			AI	Include Robin Cook's "simple" model
14	RC	49	217		AI	Investigate whether a simpler (than SS) model would be useful
15	RC	52	226		both	Modify SS so that F is explicitly modeled as an age/size effect × a year effect
16	RC	58	237		both	Rewrite SS in ADMB RE
17	RC	61			both	Distinguish between Bayesian priors and penalty functions
18	RC	65	259		both	Include calculation of DIC in SS
19	RC	67			both	Investigate whether distribution of length at age in SS implies incremental growth
20	RC	70	292		both	Investigate all data more thoroughly
21	RC	71			both	Consider alternative models with different assumptions about errors in the data
22	RC	88			BS	Use random-walk,/time-varying selectivity
23	RC	92			BS	Consider allowing time variability in catchability carefully
24	RC	99			BS	Model time variability in selectivity as a random walk
25	RC	104			BS	Model time variability in selectivity as a random walk
26	RC	118			BS	Do not estimate both M and catchability internally if catchability varies with time
27	RC	123			BS	Fix M as an age-dependent vector, using Lorenzen relationship
28	RC	134			BS	Estimate catchability internally
29	RC	145			BS	Include the longline surveys and slope trawl survey
30	RC	155			BS	Investigate whether including both longline surveys overweights large fish
31	RC	167			BS	Test weighting sensitivity by assuming agecomp is lognormal
32	RC	176			BS	Investigate large gradient problem by using alternative minimizers
33	RC	184			AI	Do not use random effects for a survey with less than annual data
34	RC	188			AI	Use the post-1986 AI trawl survey time series
35	RC	192			AI	Use a parametric selectivity function
36	RC	195			AI	Investigate sensitivity of results when using parametric selectivity
37	RC	197			AI	Investigate significance of very large changes in annual catch biomass
38	RC	201			AI	Reduce variability in catch or acknowledge that survey does not reflect biomass
39	RC	206			AI	Do not use SS unless further model exploration indicates otherwise
40	RC	212			AI	Include Robin Cook's "simple" model
41	RC	249			both	Give much more thought to choice of penalty functions, especially bounds
42	RC	264			both	Investigate whether distribution of length at age in SS implies incremental growth
43	RC	272			BS	Use random-walk,/time-varying selectivity
44	RC	273			BS	Model time variability in selectivity as a random walk
45	RC	274			BS	Include the longline surveys and slope trawl survey if weighted appropriately
46	RC	277			BS	Do not estimate both M and catchability internally if catchability varies with time
47	RC	280			AI	Conduct much more model exploration
48	RC	281			AI	Explain yearly variation in AI catches using objective criteria
49	RC	284			AI	Investigate whether a simpler (than SS) model would be useful
50	RC	288			both	Rewrite SS in ADMB RE
51	RC	289			both	Treat catch data in SS as observations as opposed to parameters
52	RC	290			both	Include calculation of DIC in SS

Table 1. Line numbers on which each unique comment begins, with comment summaries (page 2 of 3).

No.	Rev.	1st	2nd	3rd	Area	Comment summary
53	RC	294			both	Use a model that includes a length-based growth projection matrix
54	RC	297			both	Consider using SAM (Nielsen and Berg, 2014)
55	RC	298			both	Consider using a truly Bayesian approach
56	NK	4	77		both	Include all data sets that meet specified criteria
57	NK	6			both	If the longline surveys are shown to be unbiased, use them
58	NK	9			BS	Consider eliminating suspect data from the IPHC survey in the BS
59	NK	10	200		both	Set multinomial input N equal to number of sampled trips
60	NK	12			both	Weight all indices internally (SS "extra SD")
61	NK	13			both	Use either Francis or harmonic mean weighting
62	NK	14			both	Use the simplest selectivity form that gives a reasonable fit
63	NK	19	300		both	Test random-walk/time-varying vs. parametric/partitioned selectivity
64	NK	26			BS	Limit time variability to recruitment and selectivity
65	NK	30			BS	Consider relationship between dome-shaped survey selectivity and M
66	NK	32	352		BS	Use "Piner plots" to identify data sources that are in conflict with estimated M
67	NK	34			BS	Explore age-specific M (e.g., using Lorenzen function)
68	NK	35	356		both	Accept dome-shaped survey selectivity only if it occurs in many models
69	NK	37	357		BS	Estimate survey selectivity at older ages freely, with sensitivity analysis
70	NK	41	380		both	Estimate catchability internally with a "fairly non-informative" prior
71	NK	42	381		both	Estimate catchability internally with "agreed bounds" (e.g, 0.5-2 $\times$ estimate)
72	NK	45	385		both	Estimate catchability of new surveys internally with non-restrictive priors
73	NK	47			BS	Develop a prior for survey catchability "that can be generally agreed"
74	NK	50			both	For all data sets, document how they meet specified criteria
75	NK	60			both	Give primary responsibility for data documentation to the collectors
76	NK	63			both	Ask the data collectors to rank their data in order of potential bias
77	NK	68			AI	Improve documentation for the AI trawl survey time series
78	NK	74			AI	Use the post-1994 AI trawl survey time series
79	NK	85			BS	Include IPHC survey as a means of stabilizing the BS assessment
80	NK	98			BS	Examine IPHC survey data more thoroughly before using
81	NK	105			BS	Include IPHC survey to provide information on larger fish
82	NK	115			BS	Investigate anomalous 1999 and 2005 IPHC indices in the BS
83	NK	124			BS	Examine IPHC survey data more thoroughly before using
84	NK	136			AI	Consider simplifying model to improve retrospective behavior
85	NK	145			AI	Investigate alternatives to double-normal selectivity
86	NK	148			AI	Examine IPHC survey data in the AI more thoroughly before using
87	NK	160			both	Weight all indices internally (SS "extra SD")
88	NK	164			both	Downweight conflicting indices objectively
89	NK	176			BS	Evaluate possible bias of NMFS longline survey since 2010
90	NK	179			BS	Examine NMFS longline survey data in the BS more thoroughly before using
91	NK	188			AI	Examine NMFS longline survey data in the AI more thoroughly before using
92	NK	195			both	Weight all indices internally (SS "extra SD")
93	NK	206			both	Use either Francis or harmonic mean weighting
94	NK	214			both	Use Dirichlet multinomial likelihood if it becomes an option in SS
95	NK	216			both	Do not use arithmetic mean weighting
96	NK	234			both	Use the simplest selectivity form that gives a reasonable fit
97	NK	236			both	Use double-normal selectivity unless there is good reason not to
98	NK	270			BS	Use time blocks only if supported by prior data examination or external data
99	NK	277			BS	Use annually varying selectivity if it fits as well as season/gear structure
100	NK	283			BS	Fix fishery selectivity above age 8
101	NK	285			BS	Changes in mixed-gear selectivity should mirror ratios of catch by gear
102	NK	293			BS	Include at least one longline survey in order to explain cryptic biomass
103	NK	298			BS	Explore length-based random walk selectivity if it becomes an option in SS
104	NK	318			both	Consider how time-varying parameters affect reference point estimation

Table 1. Line numbers on which each unique comment begins, with comment summaries (page 3 of 3).

No.	Rev.	1st	2nd	3rd	Area	Comment summary
105	NK	322			both	Allow fishery selectivity to vary fairly freely if model is not gear-structured
106	NK	324			both	Do not allow survey selectivity or catchability to vary with time
107	NK	327			both	Account for trends in survey selectivity or catchability outside the model
108	NK	335			BS	Investigate residual "runs" in models with time-varying catchability
109	NK	337			BS	Remove time variability from parameters exhibiting little variation
110	NK	340			BS	Limit time variability to recruitment and selectivity
111	NK	343			both	Consider time-varying growth if supported by data
112	NK	347			both	Rewrite SS in ADMB RE
113	NK	365			both	Examine, rank, and document all potential sources of error in catchability
114	NK	388			both	Do not worry about large gradients if the Hessian matrix can be inverted
115	NK	412			both	Do not accept models with significant retrospective bias
116	NK	420			both	Include "Ianelli squid plots" for time-varying parameters
117	NK	425			both	Use alternative plausible historical catch scenarios to test sensitivity
118	NK	429			BS	Highlight sensitivity of model results to assumed value of "steepness."
119	NK	434			both	Consider possible effects of 1989 and 1999 regime shifts
120	JM	4	70	270	both	If IPHC survey passes further investigation, use it
121	JM	6	74	272	both	If IPHC survey passes further investigation, use it
122	JM	8	92	275	both	If NMFS longline survey passes further investigation, use it
123	JM	9	93	276	both	If NMFS longline survey passes further investigation, use it
124	JM	12	119	286	both	Do not allow "too much" time-variability in selectivity or "strange" patterns
125	JM	17	146	296	BS	Investigate selectivity differences between the two longline surveys
126	JM	19	313		BS	Consider implications of time-varying growth
127	JM	22	256	329	AI	Use a simpler form of monitoring and management, involving industry and NGOs
128	JM	25	332		both	Do not include more model features than can be supported by the data
129	JM	27	107	281	both	Weight abundance indices more heavily than sizecomps
130	JM	28	335		both	Start including fishery agecomp data
131	JM	29	336		both	Downweight sizecomps, as they are only indirect measures of abundance
132	JM	31	342		both	Include "historical" retrospectives (i.e., from previous assessments)
133	JM	33	345		both	Investigate whether a simpler (than SS) model would be useful
134	JM	41	265		both	Examine IPHC survey data in both areas more thoroughly before using
135	JM	53	268		BS	Exclude 1999 and 2005 IPHC indices in the BS if appropriate
136	JM	56			BS	Include one or both longline surveys to provide information on larger fish
137	JM	96			both	Include additional data sets to increase confidence in model results
138	JM	108	282		both	Weight agecomps heavily
139	JM	109	283		both	Downweight sizecomps, as they can mislead if growth changes over time
140	JM	111	337		BS	Use external data to help SS estimate time-varying growth
141	JM	135	292		BS	Match trawl survey selectivity to longline fishery selectivity at large sizes
142	JM	152	298		both	Structure models by season if data are sufficient
143	JM	153			AI	Consider whether data are sufficient to support seasonal structure
144	JM	157	300		both	Structure models by gear if data are sufficient
145	JM	166	303		both	Allow time variability only where supported by external data
146	JM	173	307		BS	Force trawl survey selectivity to be asymptotic
147	JM	174	308		both	Do not allow "too much" time-variability in selectivity or "strange" patterns
148	JM	182			both	Occasionally compare model's estimated survey biomass to area-swept value
149	JM	200			both	Do not estimate both M and catchability internally
150	JM	208			both	Estimate catchability internally with bounds (0.5,1.5)
151	JM	210			AI	Expect AI survey catchability to be estimated less precisely than in the BS
152	JM	227			BS	Consider implications of time-varying growth
153	JM	234	316		BS	Use trawl survey estimates at ages 1-3 as an index of recruitment
154	JM	237	319		BS	Verify that the trawl survey data sometimes include age 0 fish
155	JM	243	322		BS	Investigate whether fishing mortality for recent years is overestimated
156	JM	249	325		AI	Investigate whether M was higher than usual in the late 1980s and in 2010

Table 2. Summary of reviewer comments by ToR and area (page 1 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		1. Data currently used in the assessment models	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
1.01	AI	Explain yearly variation in AI catches using objective criteria	1	0	0	1
1.02	AI	Improve documentation for the AI trawl survey time series	0	1	0	1
1.03	AI	Reduce variability in catch or acknowledge that survey does not reflect biomass	1	0	0	1
1.04	AI	Use the post-1986 AI trawl survey time series	1	0	0	1
1.05	AI	Use the post-1994 AI trawl survey time series	0	1	0	1
1.06	both	Ask the data collectors to rank their data in order of potential bias	0	1	0	1
1.07	both	For all data sets, document how they meet specified criteria	0	1	0	1
1.08	both	Give primary responsibility for data documentation to the collectors	0	1	0	1
1.09	both	Include all data sets that meet specified criteria	0	1	0	1
1.10	both	Investigate all data more thoroughly	1	0	0	1
		1a. Potential use of IPHC longline survey data	**1**	**1**	**1**	**3**
		Examine IPHC survey in both areas more closely; use if no red flags	*1*	*1*	*1*	*3*
1a.01	BS	Examine IPHC survey data more thoroughly before using	0	1	1	2
1a.02	BS	Include IPHC survey as a means of stabilizing the BS assessment	0	1	0	1
1a.03	BS	Include the IPHC longline survey in order to explain cryptic biomass	0	1	0	1
1a.04	BS	Include IPHC survey to provide information on larger fish	0	1	1	2
1a.05	BS	Include IPHC longline and slope trawl surveys if weighted appropriately	1	0	0	1
1a.06	BS	Include the IPHC longline survey and slope trawl survey	1	0	0	1
1a.07	AI	Examine IPHC survey data more thoroughly before using	0	1	1	2
1a.08	both	Examine IPHC survey data more thoroughly before using	0	1	0	1
1a.09	both	If the IPHC longline survey is shown to be unbiased, use it	0	1	0	1
1a.10	both	If IPHC survey passes further investigation, use it	0	0	1	1
		Exclude certain years from IPHC time series in the BS if appropriate	*0*	*1*	*1*	*2*
1a.11	BS	Consider eliminating suspect data from the IPHC survey in the BS	0	1	0	1
1a.12	BS	Exclude 1999 and 2005 IPHC indices in the BS if appropriate	0	0	1	1
1a.13	BS	Investigate anomalous 1999 and 2005 IPHC indices in the BS	0	1	0	1
		Other	*0*	*0*	*1*	*1*
1a.14	BS	Investigate selectivity differences between the two longline surveys	0	0	1	1
		1b. Potential use of NMFS longline survey data	**1**	**1**	**1**	**3**
		Examine NMFS longline in both areas more closely; use if no red flags	*1*	*1*	*1*	*3*
1b.01	BS	Examine NMFS longline survey data more thoroughly before using	0	1	0	1
1b.02	BS	Include the NMFS longline survey in order to explain cryptic biomass	0	1	0	1
1b.03	BS	Include the NMFS longline survey to provide information on larger fish	0	0	1	1
1b.04	BS	Include NMFS longline and slope trawl surveys if weighted appropriately	1	0	0	1
1b.05	BS	Include the NMFS longline survey and slope trawl survey	1	0	0	1
1b.06	AI	Examine NMFS longline survey data more thoroughly before using	0	1	0	1
1b.07	both	If the NMFS longline survey is shown to be unbiased, use it	0	1	0	1
1b.08	both	If NMFS longline survey passes further investigation, use it	0	0	1	1
		Other	*0*	*1*	*0*	*1*
1b.09	BS	Evaluate possible bias of NMFS longline survey since 2010	0	1	0	1

Table 2. Summary of reviewer comments by ToR and area (page 2 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2a. How should the various data sets be weighted?	**1**	**1**	**1**	**3**
		Weighting of data in general	*1*	*1*	*1*	*3*
2a.01	BS	Test weighting sensitivity by assuming agecomp is lognormal	1	0	0	1
2a.02	both	Do not use arithmetic mean weighting	0	1	0	1
2a.03	both	Downweight conflicting indices objectively	0	1	0	1
2a.04	both	Downweight sizecomps, as they are only indirect measures of abundance	0	0	1	1
2a.05	both	Downweight sizecomps, as they can mislead if growth changes over time	0	0	1	1
2a.06	both	Set multinomial input N equal to number of sampled trips	0	1	0	1
2a.07	both	Use either Francis or harmonic mean weighting	0	1	0	1
2a.08	both	Weight abundance indices more heavily than sizecomps	0	0	1	1
2a.09	both	Weight agecomps heavily	0	0	1	1
2a.10	both	Weight all indices internally (SS "extra SD")	0	1	0	1
		Weighting of IPHC and NMFS longline survey data in particular	*1*	*0*	*0*	*1*
2a.11	BS	Address issue of how to weight IPHC and NMFS longline surveys	1	0	0	1
2a.12	BS	Investigate whether including both longline surveys overweights large fish	1	0	0	1
2a.13	BS	Weight IPHC and NMFS longline surveys internally (SS "extra SD")	1	0	0	1
		2b. What form (SS "pattern") should be used for the selectivity functions?	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
2b.01	BS	Fix fishery selectivity above age 8	1	1	0	2
2b.02	BS	Use random-walk,/time-varying selectivity	1	0	0	1
2b.03	AI	Investigate alternatives to double-normal selectivity	0	1	0	1
2b.04	AI	Investigate sensitivity of results when using parametric selectivity	1	0	0	1
2b.05	AI	Use a parametric selectivity function	1	0	0	1
2b.06	both	Test random-walk/time-varying vs. parametric/partitioned selectivity	0	1	0	1
2b.07	both	Use double-normal selectivity unless there is good reason not to	0	1	0	1
2b.08	both	Use the simplest selectivity form that gives a reasonable fit	0	1	0	1
		2c. Should the models be structured with respect to season?	**0**	**1**	**1**	**2**
		Various	*0*	*1*	*1*	*2*
2c.01	BS	Use annually varying selectivity if it fits as well as season/gear structure	0	1	0	1
2c.02	AI	Consider whether data are sufficient to support seasonal structure	0	0	1	1
2c.03	both	Structure models by season if data are sufficient	0	0	1	1
		2d. Should the models be structured with respect to gear?	**0**	**0**	**1**	**1**
		General (some reviewers merged ToR 2d comments into ToR 2b and 2c)	*0*	*0*	*1*	*1*
2d.01	both	Structure models by gear if data are sufficient	0	0	1	1

Table 2. Summary of reviewer comments by ToR and area (page 3 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2e. How much time variability should be allowed, and in which parameters?	**1**	**1**	**1**	**3**
		General	*1*	*1*	*1*	*3*
2e.01	BS	Limit time variability to recruitment and selectivity	0	1	0	1
2e.02	BS	Remove time variability from parameters exhibiting little variation	0	1	0	1
2e.03	BS	Use time blocks only if supported by prior data examination or external data	0	1	0	1
2e.04	AI	Consider simplifying model to improve retrospective behavior	0	1	0	1
2e.05	AI	Do not use random effects for a survey with less than annual data	1	0	0	1
2e.06	both	Allow time variability only where supported by external data	0	0	1	1
2e.07	both	Consider how time-varying parameters affect reference point estimation	0	1	0	1
		Time-varying selectivity	*1*	*1*	*1*	*3*
2e.08	BS	Changes in mixed-gear selectivity should mirror ratios of catch by gear	0	1	0	1
2e.09	BS	Model time variability in selectivity as a random walk	1	0	0	1
2e.10	both	Account for trends in survey selectivity ... outside the model	0	1	0	1
2e.11	both	Do not allow survey selectivity ... to vary with time	0	1	0	1
2e.12	both	Do not allow "too much" time-variability in selectivity…	0	0	1	1
2e.13	both	Allow fishery selectivity to vary fairly freely if model is not gear-structured	0	1	0	1
		Time-varying catchability	*1*	*1*	*0*	*2*
2e.14	BS	Consider allowing time variability in catchability carefully	1	0	0	1
2e.15	BS	Investigate residual "runs" in models with time-varying catchability	0	1	0	1
2e.16	BS	Model time variability in catchability as a random walk	1	0	0	1
2e.17	both	Account for trends in survey ... catchability outside the model	0	1	0	1
2e.18	both	Do not allow survey ... catchability to vary with time	0	1	0	1
		Time-varying growth	*0*	*1*	*1*	*2*
2e.19	BS	Consider implications of time-varying growth	0	0	1	1
2e.20	BS	Use external data to help SS estimate time-varying growth	0	0	1	1
2e.21	both	Consider time-varying growth if supported by data	0	1	0	1
		2f. What constraints, if any, should be placed on old-age survey selectivity?	**0**	**1**	**1**	**2**
		Various	*0*	*1*	*1*	*2*
2f.01	BS	Consider relationship between dome-shaped survey selectivity and M	0	1	0	1
2f.02	BS	Estimate survey selectivity at older ages freely, with sensitivity analysis	0	1	0	1
2f.03	BS	Force trawl survey selectivity to be asymptotic	0	0	1	1
2f.04	BS	Match trawl survey selectivity to longline fishery selectivity at large sizes	0	0	1	1
2f.05	both	Accept dome-shaped survey selectivity only if it occurs in many models	0	1	0	1
2f.06	both	Do not allow … "strange" patterns	0	0	1	1
		2g. What constraints, if any, should be placed on survey catchability?	**1**	**1**	**1**	**3**
		Use a prior distribution for catchability	*0*	*1*	*1*	*2*
2g.01	BS	Develop a prior for survey catchability "that can be generally agreed"	0	1	0	1
2g.02	AI	Expect AI survey catchability to be estimated less precisely than in the BS	0	0	1	1
2g.03	both	Estimate catchability internally with a "fairly non-informative" prior	0	1	0	1
2g.04	both	Estimate catchability of new surveys internally with non-restrictive priors	0	1	0	1
		Adjust bounds on catchability	*0*	*1*	*1*	*2*
2g.05	both	Estimate catchability internally with "agreed bounds" (e.g, 0.5-2 ´ estimate)	0	1	0	1
2g.06	both	Estimate catchability internally with bounds (0.5,1.5)	0	0	1	1
		Other	*1*	*1*	*1*	*3*
2g.07	BS	Estimate catchability internally	1	0	0	1
2g.08	both	Examine, rank, and document all potential sources of error in catchability	0	1	0	1
2g.09	both	Occasionally compare model's estimated survey biomass to area-swept value	0	0	1	1
		2h. How should big gradients be dealt with in apparently converged models?	**1**	**1**	**0**	**2**
		Various	*1*	*1*	*0*	*2*
2h.01	BS	Investigate large gradient problem by using alternative minimizers	1	0	0	1
2h.02	both	Do not worry about large gradients if the Hessian matrix can be inverted	0	1	0	1

Table 2. Summary of reviewer comments by ToR and area (page 4 of 4).

No.	Area	**Heading**/*subheading*/comment summary	RC	NK	JM	Sum
		2i. Other comments (reviewers' choice)	**1**	**1**	**1**	**3**
		General modeling philosophy	*1*	*1*	*1*	*3*
2i.01	AI	Conduct much more model exploration	1	0	0	1
2i.02	AI	Do not use SS unless further model exploration indicates otherwise	1	0	0	1
2i.03	both	Do not accept models with significant retrospective bias	0	1	0	1
2i.04	both	Do not include more model features than can be supported by the data	0	0	1	1
		Natural mortality rate	*1*	*1*	*1*	*3*
2i.05	BS	Do not estimate both M and catchability internally if catchability varies with time	1	0	0	1
2i.06	BS	Explore age-specific M (e.g., using Lorenzen function)	0	1	0	1
2i.07	BS	Fix M as an age-dependent vector	1	0	0	1
2i.08	BS	Fix M as an age-dependent vector, using Lorenzen relationship	1	0	0	1
2i.09	BS	Use "Piner plots" to identify data sources that are in conflict with estimated M	0	1	0	1
2i.10	AI	Investigate whether M was higher than usual in the late 1980s and in 2010	0	0	1	1
2i.11	both	Do not estimate both M and catchability internally	0	0	1	1
		Investigation of suspect results or model features	*1*	*0*	*1*	*2*
2i.12	BS	Investigate whether fishing mortality for recent years is overestimated	0	0	1	1
2i.13	BS	Verify that the trawl survey data sometimes include age 0 fish	0	0	1	1
2i.14	AI	Investigate significance of very large changes in annual catch biomass	1	0	0	1
2i.15	both	Investigate whether distribution of length at age in SS implies incremental growth	1	0	0	1
		Alternative models ("simple")	*1*	*0*	*1*	*2*
2i.16	AI	Include Robin Cook's "simple" model	1	0	0	1
2i.17	AI	Investigate whether a simpler (than SS) model would be useful	1	0	0	1
2i.18	AI	Use a simpler form of monitoring and management, involving industry and NGOs	0	0	1	1
2i.19	both	Investigate whether a simpler (than SS) model would be useful	0	0	1	1
		Retrospective diagnostics	*0*	*1*	*1*	*2*
2i.20	both	Include "historical" retrospectives (i.e., from previous assessments)	0	0	1	1
2i.21	both	Include "Ianelli squid plots" for time-varying parameters	0	1	0	1
		Features to add in SS and use once added	*1*	*1*	*0*	*2*
2i.22	BS	Explore length-based random walk selectivity if it becomes an option in SS	0	1	0	1
2i.23	both	Include calculation of DIC in SS	1	0	0	1
2i.24	both	Modify SS so that F is explicitly modeled as an age/size effect ´ a year effect	1	0	0	1
2i.25	both	Rewrite SS in ADMB RE	1	1	0	2
2i.26	both	Treat catch data in SS as observations as opposed to parameters	1	0	0	1
2i.27	both	Use Dirichlet multinomial likelihood if it becomes an option in SS	0	1	0	1
		Sensitivity testing	*0*	*1*	*0*	*1*
2i.28	BS	Highlight sensitivity of model results to assumed value of "steepness."	0	1	0	1
2i.29	both	Consider possible effects of 1989 and 1999 regime shifts	0	1	0	1
2i.30	both	Use alternative plausible historical catch scenarios to test sensitivity	0	1	0	1
		Alternative models (fully age- or length-structured)	*1*	*0*	*0*	*1*
2i.31	both	Consider alternative models with different assumptions about errors in the data	1	0	0	1
2i.32	both	Consider using a truly Bayesian approach	1	0	0	1
2i.33	both	Consider using SAM (Nielsen and Berg, 2014)	1	0	0	1
2i.34	both	Use a model that includes a length-based growth projection matrix	1	0	0	1
		Prior distributions and penalty functions	*1*	*0*	*0*	1
2i.35	both	Distinguish between Bayesian priors and penalty functions	1	0	0	1
2i.36	both	Give much more thought to choice of penalty functions, especially bounds	1	0	0	1
		Use of additional data (other than longline surveys)	*0*	*0*	*1*	*1*
2i.37	BS	Use trawl survey estimates at ages 1-3 as an index of recruitment	0	0	1	1
2i.38	both	Include additional data sets to increase confidence in model results	0	0	1	1
2i.39	both	Start including fishery agecomp data	0	0	1	1

Attachment 1: Terms of Reference

1. Evaluate and provide recommendations on data used in the assessment models. In particular:
    a. Should data from the IPHC longline survey be used in either assessment?
    b. Should data from the NMFS longline survey be used in either assessment?
2. Evaluate and provide recommendations on model structure, assumptions, and estimation procedures. In particular:
    a. How should the various data sets be weighted?
    b. What form (i.e., Stock Synthesis "pattern") should be used for the selectivity functions?
    c. Should the models be structured with respect to season?
    d. Should the models be structured with respect to gear type?
    e. How much time variability should be allowed, and in which parameters?
    f. What constraints, if any, should be placed on survey selectivity at older ages?
    g. What constraints, if any, should be placed on survey catchability?
    h. How should large gradients be dealt with in otherwise apparently converged models?
    i. Anything else on which the reviewers care to comment.

Attachment 2: Agenda

CIE Review of the EBS and AI Pacific cod stock assessment models

Alaska Fisheries Science Center
7600 Sand Point Way NE, Seattle, WA 98115
February 16-19, 2016
Building 4; Room 2039 (except Wednesday afternoon), Room 2143 (Wednesday afternoon)

Review panel chair: Anne Hollowed, Anne.Hollowed@noaa.gov
Senior assessment author: Grant Thompson, Grant.Thompson@noaa.gov
Security and check-in: Sandra Lowe, Sandra.Lowe@noaa.gov (206)526-4230

Sessions will run from 9 a.m. to 5 p.m. each day, with time for lunch and morning and afternoon breaks. Discussion will be open to everyone, with priority given to the panel and senior assessment author.

Tuesday, February 16

Preliminaries:
- 09:00 Introductions and adoption of agenda—Anne

Data sources (current and potential):
- 09:10 Overview of data types used in the assessments—Grant
- 09:20 Observer program—Craig Faunce, AFSC FMA Division
- 09:50 Catch accounting system and in-season management—Mary Furuness, AKRO SF Division
- 10:20 Break
- 10:30 EBS trawl survey—Bob Lauth, AFSC RACE Division
- 11:00 AI trawl survey—Wayne Palsson, AFSC RACE Division
- 11:30 IPHC longline survey—Anna Henry, IPHC
- 12:00 Lunch
- 13:00 NMFS longline survey—Dana Hanselman (via WebEx)
- 13:30 Ageing—Tom Helser, AFSC REFM Division

Assessment models:
- 14:00 Assessment history—Grant
- 15:00 Break
- 15:10 Current assessments—Grant
- 16:10 Discussion—Everyone
- 16:40 Assignments for models to be presented on Wednesday—Panel

Wednesday, February 17

Review of models assigned on Tuesday—Grant
Discussion, real-time model runs—Everyone
Assignments for models to be presented on Thursday—Panel

Thursday, February 18

Review of models assigned on Wednesday—Grant
Discussion, real-time model runs—Everyone
Assignments for models to be presented on Friday—Panel

Friday, February 19

Review of models assigned on Thursday—Grant
Discussion, real-time model runs—Everyone
Report writing (time permitting)—Panel

Attachment 3: CIE reviewer reports (main substantive text only)

Reviewer 1: Robin Cook

Executive summary

Bering Sea

Overall, the change from parametric descriptions of selectivity to random effects varying over time are likely to be important improvements to the Bering Sea model that reflect recent trends in stock assessment modelling and should be pursued. There may be advantages in modelling time varying parameters as a random walk to exploit "memory" in model.

Estimating age specific selectivity with annual variability for older fish is likely to be affected by large errors associated with small samples. I would suggest collapsing fish older than 10-12 into a plus group and setting the selectivity of the plus group to be equal the oldest true age.

While it may be possible to estimate fixed M and fixed Q given adequate contrast in the data, allowing Q to vary over time could mean that model mis-specification may arbitrarily emerge as variations in annual changes to Q. It may be better to fix M as an age dependent vector. Given that survey sampling protocols seek to minimise random changes, it is probably better to model annual changes in Q as a random walk to avoid over-fitting the data.

It is probably more useful to estimate Q within the model and regard it as a value that reconciles the assessment scale to the survey scale. Fixing Q within the model will add a degree of rigidity that may lead to severely biased estimates of fishing mortality, especially where the catch is treated as an exact constant.

A number of the exploratory runs performed at the meeting considered the inclusion of the IPHC longline, NMFS longline, and NMFS slope trawl survey in addition to the conventional use of the NMFS shelf trawl survey. Unlike the shelf survey, the three other surveys do not sample smaller fish and the length frequencies of their samples resembles more closely the commercial fishery length compositions. Trial runs for the EBS when these additional surveys were included tended to result in higher ending F and lower ending biomass so their inclusion is pertinent to management decisions. There is an issue about the appropriate weighting to give these surveys and this needs further exploration. Weighting the data by their estimated precision would seem appropriate.

An exploratory run that used Francis weighting for length and age compositions resulted in much lower estimates of F and much higher values of ending biomass. The sensitivity to an alternative weighting method is a cause for concern. An alternative error structure might be explored as a sensitivity test by fitting the model to the index of number-at-age in the survey which assumes that the observed number at age is lognormally distributed with age specific error distributions.

Aleutian Islands

This assessment differs from the Bering Sea assessment in that recorded catch biomass shows very large inter-annual changes and the trawl survey data series is not continuous. There are many missing years and there have been changes to the survey protocol over time. The base model used in exploratory runs (15.7) showed poor retrospective properties and a high sensitivity to the data used in the trawl survey.

At present it seems preferable to use parametric functions for the fishery and survey selectivity and to fix these over time. However, the very large range of outcomes from exploratory runs is a concern in the assessment and further analysis is required.

The very large annual variations in estimated F which may change by a factor of 2 or more in a single year, does not appear to produce a response in the stock biomass and implies recruitment is matching the change in biomass. This seems highly unlikely and suggests a problem with the assessment. Either the catches are less variable than they appear or the survey index does not adequately reflect real changes to stock biomass.

In the past a Tier 5 model has been used for the assessment of AI Pacific cod and in view of the problems with the Tier 3 model a simpler modelling approach is desirable as a fallback assessment. It might be possible to go further than simply smoothing the biomass indices by using a model similar to that outlined in Annex A, Table A2. Whether a simpler modelling approach is useful should be investigated if only to understand better which data contain useful information.

Other comments

Modifying the SS3 projection equation so that fishing mortality is explicitly modelled as the product of an age (or size) effect and a year effect offers scope for removing errors in the catches (provided these are treated as observations rather than constants), as well as exploring alternative models of fishing mortality (e.g. fixed effects, random walk) or even using fishing effort as a covariate.

Estimating the random effects in a conventional likelihood framework within the core ADMB requires external intervention to tune the analysis. Where complex random effects models are assumed, as in the case of some selectivity models, there remains some doubt as to the reliability of the estimates. An obvious way forward would be to redevelop Synthesis within the RE version of ADMB so that the random effects can be estimated within a conventional and tested mathematical framework.

The distinction between “Bayesian priors” and “penalty functions” as used by SS3 needs to be made clearer particularly where “uniform priors” are used. Where the latter represent bounds on the parameters, this should be made explicit especially where the converged model solution lies on a bound.

Where random effects are used, it was difficult to judge model performance in relation to goodness of fit versus the number of parameters. It would be desirable to calculate the Deviance Information Criterion (DIC) to try to overcome this problem so that different models can be compared.

It is worth investigating whether the assumption of normal distributions of length at age in SS3 is consistent with strictly incremental growth. It is possible that the model structure implies some fish much decline in length to be consistent with the size distribution model.

While Synthesis is an important and effective tool in the assessor’s kit, it would be worth devoting effort to a more thorough investigation of the various data components before applying Synthesis as well as giving thought to alternative models that make different assumptions about errors in the data.

Bering Sea

At the review, the principal innovations proposed for a new SS3 configuration were:

1. Each year consisted of a single season instead of five.
2. A single fishery was defined instead of nine season-and-gear-specific fisheries.

3. Selectivity for both the fishery and survey was modeled using a random walk with respect to age (SS selectivity-at-age pattern #17) instead of the usual double normal.
4. Selectivity for both the fishery and the survey were allowed to vary annually.
5. Survey catchability was allowed to vary annually.
6. Initial abundances were estimated for the first ten age groups instead of the first three.
7. The natural mortality rate was estimated internally.
8. The base value of survey catchability was estimated internally.

Points 1-2 greatly simplify the model and avoid the need to estimate a large number of selectivity parameters. Point 3 is an important change that allows the data to determine the shape of fishery and survey selectivity. Allowing these to vary over time (point 4) then allows selectivity to evolve as the fishery changes. This is a natural way to accommodate changes to the activity and developments in different fleets that target different age and size components of the stock. Overall, these are likely to be important improvements to the model that reflect recent trends in stock assessment modelling and should be pursued.

While the use of random effects models has advantages in terms of the number of parameters to be estimated and model flexibility, there is danger of allowing too much flexibility. The annual changes to catchability (points 4 and 5) need to be considered carefully. My understanding of model 15.6, which formed the basis of most of the CIE requested runs, is that annual changes to selectivity and survey catchability were independent with respect to time. The danger of such an approach is that annual changes may simply reflect noise in the data rather than any true signal, because the model has no "memory" of what happened in the previous year. In many fisheries fleet behavior does not change substantially from year to year but evolves gradually over time. Thus, unless there is a "shock" to the system, one would expect selectivity in successive years to be correlated. Such correlation should be exploited in the model by, for example, modeling selectivity (sel) as:

$$sel(age,year)=sel(age, year-1)*\exp(e_{year})$$

where e_{year} is a random innovation drawn from a normal distribution.

Using a random walk with respect to age to model selectivity is a fair enough assumption but is perhaps not strictly necessary if selectivity is modelled as a random walk over time. I would suggest estimating age dependent selectivities in the initial year as free parameters and then allow the base selectivity pattern to follow a random walk over time. Because the model remembers the previous year's selectivity, it means that all the data inform the estimates of fleet selectivity. This is an increasingly common assumption in current stock assessment models (e.g. Nielsen and Berg, 2014; Cook et al., 2015).

As regards the age range for estimating selectivity, it would seem desirable to consider only those older age groups that are adequately sampled. Unfortunately, misspecification of selectivity on older age groups that are poorly sampled can have a major effect on the assessment. If the estimates of fishing mortality obtained in the runs shown in Figure 1 are approximately correct, and if M is around 0.34, this implies values of total mortality in the region of 0.8. With such an exploitation rate, fish age 10 or older are likely to be very rare in samples. Estimating age specific selectivity with annual variability for older fish is therefore likely to be affected by sampling error. I would suggest collapsing fish older than 10-12 into a plus group and setting the selectivity of the plus group to be equal the oldest true age.

It is noteworthy that model 15.6 is configured to estimate both natural mortality (M) and base survey catchability (Q), which are typically highly correlated and difficult to estimate jointly. Estimating both these parameters and allowing survey catchability to change with little constraint over time seems imprudent. While it may be possible to estimate fixed M and fixed Q given adequate contrast in the data,

allowing Q to vary over time demands a great deal of the data, especially when the assumption of a fixed M over time and age is clearly very unrealistic. It means that model mis-specification may arbitrarily emerge as variations in annual changes to Q. It may be better to fix M as an age dependent vector, determined for example by the Lorenzen relationship as shown in Annex A (Fig A.1), which used mean weight at age to estimate M. In this case, the estimated value of M over the mid-upper age classes is not very dissimilar to the value of 0.34 often used for this stock. Adopting an age or size dependent M value may be relevant to the assumptions used in these assessments for dome shaped selectivity. With M determined, it is then possible to estimate Q albeit conditioned on the assumed value of natural mortality. Given that survey sampling protocols seek to minimise random changes, it is probably better to model annual changes in Q as a random walk to avoid over-fitting the data.

There is much discussion in the Region of the value of survey catchability. There is clearly a desire for swept area estimates of Q to be seen to be close to values estimated from stock assessment models. It is undoubtedly of interest to compare such estimates and to try to understand the causes of any differences. However, there should be no surprise if such estimates differ and it is a mistake to force a given value of Q into the assessment since the assumptions on which the calculations are based are quite different. Where Q is estimated experimentally there is an assumption that the survey is sampling the same population as the fishery, and that the scaling factors used to raise trawl survey samples to absolute abundance are both accurate and unbiased. None of these assumptions is completely correct and there will be considerable uncertainty surrounding them. It is probably more useful to estimate Q within the model and regard it as a value that reconciles the assessment scale to the survey scale. Fixing Q within the model will add a degree of rigidity that may lead to severely biased estimates of fishing mortality, especially where the catch is treated as a known constant.

A number of the exploratory runs performed at the meeting considered the inclusion of the IPHC longline, NMFS longline and NMFS slope trawl survey in addition to the conventional use of the NMFS shelf trawl survey. It is generally considered good practice to include all the available data unless there are strong reasons to omit it. On that basis, all the surveys should be included. Unlike the shelf survey, the three other surveys do not sample smaller fish and the length frequencies of their samples resembles more closely the commercial fishery length compositions. It would appear therefore that these other surveys sample a part of the population that is not so well sampled by the shelf survey. Trial runs for the EBS when these additional surveys were included tended to result in higher ending F and lower ending biomass so their inclusion is pertinent to management decisions. There is an issue about the appropriate weighting to give these surveys. One run (BS_Model_15pt6_C_extraSD) weighted the indices by their respective standard deviation estimated internally, and this reduced the higher estimates of F seen in the other runs. Weighting the data by their estimated precision would seem appropriate and may prove the best way forward. There is, however, a somewhat different issue which is that these additional surveys all appear to sample a similar size range of the population and adding three similar surveys may bias the assessment toward the population seen by these surveys. Some exploration of this issue is required, but as mentioned, weighting by the precision of the data may be the appropriate solution.

A central feature of SS3 is that length and age compositions are fit as proportions rather than numbers at length or age. This requires an estimate of the effective sample size which is generally much lower than the actual number of fish sampled. In model runs carried out before and during the review sample sizes were constrained to be in the region of 300. An exploratory run (15.6_Francis, Figure 1) used an alternative weighting that resulted in much lower estimates of F and much higher values of ending biomass. The sensitivity to an alternative weighting method is a cause for concern though it should be noted that in the guidance notes for the use of Francis weighting there is a caution that “The large number of options available in SS makes it very difficult to be sure that what this function does is appropriate for all combinations of options”. There is no simple answer to this issue but something that might be explored as a sensitivity test would be to fit the model to the index of number-at-age in the survey as

described in Annex A, Table A1. This assumes that the observed number at age is lognormally distributed with age specific error distributions. As an alternative error structure, it is not without its own problems (e.g. correlated errors), but if the data are rich in abundance information it would offer an insight into the robustness of the assessment.

Term of reference 2h requests advice on models that have apparently converged yet with a large gradient at the minimum. Where the parameter covariance matrix is calculable, there is some reassurance that a meaningful minimum has been reached. However, this is a technical problem and its resolution will depend on the algorithm used to minimize the negative log-likelihood. If SS3 offers a choice of minimization routines these could be explored to try to diagnose the problem.

Aleutian Islands

In principle much of the discussion relating to the EBS assessments should apply to the Aleutian Islands. However, there are at least two important differences that need to be considered. Firstly, the recorded catch biomass shows very large inter-annual changes that are apparently related to major changes in fishery management from year to year. Secondly, the trawl survey data series is not continuous. There are many missing years and there have been changes to the survey protocol over time. The latter point is of particular relevance since surveys are an important point of reference for an assessment and with missing or inadequately standardized data the use of random effects in the model may be unwise. The base model used in exploratory runs (15.7) showed poor retrospective properties and when the earliest two years of survey data were omitted, a radically different estimate of ending biomass and fishing mortality was observed (Figure 2, models 15.7 and 15.8), suggesting a high sensitivity to the trawl survey. While there was a significant change in the survey between 1994 and 1997 in terms of tow duration this does not seem sufficient to justify the removal of earlier years from the analysis. It also shortens the time series of available data substantially for an assessment already lacking in calibration data.

When selectivity was modelled with parametric functions fixed over time, the retrospective pattern improved which suggests a more rigid model may be better when calibration data are scarce or unreliable. At present, therefore, it seems preferable to use parametric functions for the fishery and survey and to fix these over time, but the very large range of outcomes shown in Figure 2 does not inspire confidence in the assessment and further analysis is required.

One issue that needs investigation is the significance of the very large changes in the annual catch biomass. Since the catch is treated by SS3 as a known parameter, the variability in the catches is translated directly into variability in estimates of fishing mortality because the assessment suggests stock biomass only shows very gradual change (see for example Annex A, Figure A3 that show F from model 15.7). Given the very large annual variations in estimated F which may change by a factor of 2 or more in a single year, one might expect to see a response in the stock biomass, but this is not apparent and implies recruitment is matching the change in biomass. This seems highly unlikely and suggests a problem with the assessment. Either the catches are less variable than they appear or the survey index does not adequately reflect real changes to stock biomass.

For the reasons above, I did not feel that the SS3 models were currently in a state to form the basis of an assessment through further model exploration, especially those using additional surveys may yet prove adequate.

In the past a Tier 5 model has been used for the assessment of AI Pacific cod. This model simply smooths the IPHC longline survey and trawl survey indices using a random walk. Such methods can be of use especially if the indices are a true reflection of the biomass trend, but inevitably offer little insight into the stock dynamics. It might be possible to go one step further using a model similar to that outlined in

Annex A, Table A2. That model assumed that fishing mortality follows a random walk which may be too strong an assumption if management intervention has introduced a series of shocks to the fishery. The random walk assumption could be relaxed so that fishing mortality was modelled as a purely random effect with a large standard deviation.

Whether a simpler modelling approach is useful should be investigated if only to understand better which data contain useful information.

Other comments

The population projection model within Synthesis appears to treat the observed catches as parameters rather than as observations. Where catches are very precise this approach may work well, but it is a strong assumption and for many stocks for which Synthesis is used the assumption of exact catches is hard to justify. In the case of EBS Pacific cod there are good reasons to suppose that recorded catches are precisely known for recent years, but historically this is probably not the case. Furthermore, little appears to be known about catches outside the US EEZ, which even in recent years may have an impact on the stock. A model that avoids the need to treat catches as known is therefore highly desirable. Modifying the projection equation so that fishing mortality is explicitly modelled as the product of an age (or size) effect and a year effect offers much more scope for removing errors in the catches by treating these as observations rather than constants. It also allows exploring alternative models of fishing mortality (e.g. fixed effects, random walk) or even using fishing effort as a covariate.

Until fairly recently, selectivity in the EBS assessment was modelled using parametric functions. Time varying selectivity was handled by dividing the time series into blocks where each block has its own selectivity values. In the models discussed at the review selectivity was sometimes modelled with random effects both over age and year. Estimating the random effects in a conventional likelihood framework within the core ADMB requires external intervention to tune the analysis. Where complex random effects models are assumed as in the case of some selectivity models there remains some doubt as to the reliability of the estimates. An obvious way forward would be to redevelop Synthesis within the RE version of ADMB so that the random effects can be estimated within a conventional and tested mathematical framework. This would avoid the need for ad hoc tuning and potentially would speed up the assessment process.

In configuring Synthesis some model parameters are constrained by penalty functions that are added to the likelihood and informally referred to as “priors”. These of course are not priors in a Bayesian sense and some care is required in their interpretation as a result. Some model parameters are described as having “uniform” priors implying no constraints on the parameters when in practice bounds are set to prevent estimates reaching values considered unrealistic and are therefore highly informative. This differs substantially from a Bayesian uniform prior where bounds are set primarily to avoid the MCMC chain sampling values outside the posterior distribution rather than setting limits on acceptable parameter values. Such a prior is uninformative and will give true unconstrained maximum likelihood estimates of the parameters. Given that the penalty function may be influential in parameter estimates in Synthesis much more thought needs to be given to the choice of these functions and more attention paid to their influence in the estimates especially where bounds are reached. Hitting a bound would tend to suggest insufficient information in the data to estimate the parameters.

SS3 allows a very large range of models to be fit to the data. This often means that the number of parameters being fit varies greatly as the assessment is developed. As is well known, more parameters usually mean a better fit to the data, but not necessarily a better model. In a likelihood approach model performance can be evaluated using the AIC which trades model fit against the number of parameters. In the review carried out at this meeting, it was difficult to compare models using AIC as the number of

effective parameters is not clear in random effects models. As a result, it was difficult to judge model performance in relation to goodness of fit versus the number of parameters. It would be desirable to calculate the Deviance Information Criterion (DIC) to try to overcome this problem so that it is clear when a model is overparameterised.

A particularly important feature of SS3 is that it can make use of both length and age data. The underlying model, however, is age-based and the population length composition is reconstructed from the dispersion around the mean length at age. This may well be an adequate assumption but one issue that perhaps merits investigation is whether this approach implies non-incremental growth since the assumption of strictly normal length distributions may not be compatible with the requirement that individual fish cannot get smaller as they age (except in exceptional circumstances). The question is whether, given a normal distribution at age a, the distribution of the same year class at age a+1 is simply a normal distribution centred on the mean length at age or some other distribution? Where length data drive the assessment inconsistency with incremental growth may lead to bias.

Conclusions and recommendations

For the Eastern Bering Sea SS3 models that use random effects to model selectivity by year, and perhaps age, are the preferred configurations at this stage of assessment development. Time varying parameters may be better modeled using a random walk to prevent over-fitting the data. The longline surveys and slope trawl survey should be included provided an appropriate way of weighting the data can be found. This needs to take into account the survey sampling precision as well as weighting relative to the shelf survey to avoid over-emphasis on the deeper water component of the stock. It is probably better to fix natural mortality externally if survey catchability is estimated internally, and especially if survey catchability is treated as a time varying parameter.

Much more model exploration is required for the Aleutian Islands assessment. It is especially desirable to try to explain on the basis of objective criteria why the historical catch shows such large inter-annual variability. The relative scarcity and lower reliability of fishery independent data to calibrate the assessment also makes the current Tier 3 models rather uncertain. While developing the Tier 3 model, consideration should also be given to enhancing the Tier 5 model to include a simple population model in order to obtain a little more information from the data as opposed to simply smoothing the time series.

SS3 is a well established and powerful tool that can be used both for data exploration and full assessments. A number of aspects of the tool deserve consideration for the future development of the model. This includes a more formal way of estimating random effects through ADMB RE, treating the catch data as observations as opposed to parameters, and providing a statistic such as the DIC to compare best models when random effects are being used.

Stock Synthesis appears to be the only modelling tool considered when a full population dynamic model is fitted. While it is an important and effective tool in the assessor's kit, it would be worth devoting effort to a more thorough investigation of the various data components before applying Synthesis, as well as giving thought to alternative models that make different assumptions about the data. For example, given the major presence of length frequency data, a model that used a length based projection matrix might offer useful insights into the information contained in the data and treat growth in a more realistic fashion. Similarly models such as SAM (Nielsen and Berg, 2014), used by ICES, might provide a contrast to the multinomial assumption implemented in SS3. Truly Bayesian approaches that provided true estimates of the parameter posterior distributions may be more informative about the data than the application of penalty functions.

Reviewer 2: Neil Klaer

Executive summary

Generally, if a data source provides useful information, can fit within an assessment model structure, has been shown to be reliably collected and standardized, and is likely to be unbiased or bias can be accounted for, then it should be included in the stock assessment model. I have recommended that the IPHC and NMFS longline surveys and associated composition data be included in both the EBS and AI models conditional on documentation that examines those surveys for potential bias regarding Pacific cod. The IPHC in the EBS in particular may require truncation to eliminate suspect point estimates.

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.

For relative weighting of various data sets, I recommend estimation of an additional sd for all abundance indices, and either the Francis or harmonic mean weighting procedures for composition data.

Options for selectivity patterns are primarily among simple logistic and double-normal by size, and random walk by age. The simplest pattern that allows reasonable model fit to available composition data should be used. The most complex random walk by age pattern is most suited for application to combined fisheries composed of differing gear types, although there may be a question about the implementation of it in SS regarding large final gradients.

Allowing time-varying selectivity that is a random walk by age annually for a fishery with multiple gear types is an innovation that I have not seen previously. As many current SS assessments grapple with highly partitioned fishery data, such a procedure has the potential for wide application. I 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.

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 appears to exist for recruitment and fishery selectivity in the EBS model.

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 selectivity 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 EBS 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.

Models that estimate the shelf bottom trawl survey q using a fairly non-informative prior (as in EBS 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. Work should be commenced on the development of a prior distribution for EBS shelf bottom trawl survey q that can be generally agreed.

Evaluate and provide recommendations on data used in the assessment models (general)

As a general principle, we all understand that data to be potentially included in a stock assessment model first need an examination to determine whether they measure important aspects of stock dynamics that can be included in a stock assessment model, are collected and standardized in a rigorous manner, and are likely to be unbiased or any bias has been measured and can be accounted for. Ideally, this examination for each separate input data set would be well documented, updated as required, and provided as support information for any stock assessments. Most stock assessments do not reach this ideal. For the EBS and AI assessments, such data documentation specifically for stock assessment support does not exist. However, during the review, presentations were made that described data collection methodologies and the process used to prepare the data for use in stock assessments which could form the basis for such documentation.

The most difficult input data question regards possible bias. Normally, it is the data collectors who have the most information about changes in collection procedures, unexpected changes in data signals, potential for non-representative sampling and the like. Input data documentation should include accounts by the data collectors on these aspects, and the potential bias that may have been introduced. Where several data sources provide similar information (e.g. alternative survey abundance indices with similar gear selection), it may also be useful to ask data collectors to rank the alternatives according to potential bias. Such information may then be used by stock assessment authors when preferentially weighting various data sets.

A particular example examined during the review that illustrates the usefulness of improved documentation was for the AI trawl survey abundance index and associated composition data. A list of 10 historical changes in survey design was provided, but it was acknowledged that the input data had not been subjected to a detailed examination regarding those changes to potentially quantify their effects. As some changes appeared to be substantial but also open to desk-top investigation (e.g. any apparent shift in selectivity pattern due to the change from 30 to 15 minute trawls), my initial reaction was to not use the series trend until appropriate investigations had been made. Subsequent discussions concluded, with the help of data collectors, that changes since 1997 in survey methodology were unlikely to have caused substantial bias in the index, so it was agreed that the index was usable from that year forward.

Generally, if a data source provides useful information, can fit within an assessment model structure, has been shown to be reliably collected and standardized, and is likely to be unbiased or bias can be accounted for, then it should be included in the stock assessment model.

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

Bering Sea

The shelf trawl survey in most/all EBS models appears to require dome selectivity in comparison with the fisheries regardless of whether the fisheries are highly partitioned according to gear and season, or

selectivity is allowed to change through time (e.g. both models 11.5 and 15.6). Ideally, abundance index size/age selection would be reflective of the population – i.e. asymptotic at a low age. Models that include the trawl survey alone have considerable flexibility to alter abundance trends for older age classes not well indexed that may have a heavy influence on population SSB trends. There is an advantage therefore, to include index information for those older age classes if such indices exist. In this case, candidates are the IPHC and NMFS longline and the slope trawl surveys.

In all cases (and IPHC in particular), the available additional surveys were primarily designed to index species other than Pacific cod. Desk-top studies of the suitability of application of these surveys as potential indices of abundance to Pacific cod in particular are currently unavailable, so judgment of whether to include them into an active assessment model is only evaluated here based on presentations of survey procedures during the review, general comparisons among available indices, and the apparent performance of models that include various index combinations.

The IPHC primary objectives are to provide CPUE, length and age composition, information on abundance distributional changes for juveniles and adults for Pacific halibut. Secondary objectives are to provide information on bycatch species and a platform for specialized projects. We learned through presentations that a number of factors (different hook size to commercial Pacific cod fishing, first 20 hooks per skate sampled for bycatch, bait used, areas sampled) may not be optimal for Pacific cod, but Pacific cod are the most-often encountered bycatch species by the survey (at least in Areas 4A, 4B, 4C and 4D – covering the EBS and AI regions). This suggests that IPHC survey trends at least require examination, and that there are no reasons yet identified that imply an index bias, just sources of possibly random measurement error.

The aggregated size composition from the IPHC survey indicates a selected size range well to the right of the shelf trawl survey in the EBS, and slightly to the right of the longline fishery, NMFS longline survey and slope trawl survey (Figure 1). This indicates that the IPHC index can potentially provide useful abundance information for the older age-classes that are not indexed by the shelf trawl survey if that survey selectivity is dome-shaped.

A comparison of general index trends in the EBS (Figure 2) does not show a lot of consistency among available indices, although the different selectivity associated with those indices makes interpretation more difficult. The IPHC survey seems to exhibit trends that are least consistent with the other available indices. A shift of the IPHC survey several years to the left shows perhaps some consistency with the trawl survey. Biologically, it is not possible for the true abundance of older year classes in the Pacific cod population to change radically from one year to the next. There are two substantial drops in the IPHC index that seem biologically implausible – in 1999 and 2005. Further work is needed to investigate the cause of these changes in particular, and whether the index requires refinement in application to Pacific cod.

Among the meeting requests were those that included various new index combinations to be added to the EBS model, while also estimating an additional sd. The additional sd accounts for apparent error that is required to be added to an index for the model to be balanced, given the information from all other data sources in the model (model 15.6 extra sd). That model adds a large sd value to the IPHC index, mostly to better account for the apparent error in the 1999 index value.

Before deciding to include the IPHC longline index and associated lengths in a proposed central EBS SS model, an investigation into the properties of the EBS IPHC longline index in relation to Pacific cod in particular should be done. The investigation should examine the 1999 and 2005 points especially to see if justification exists for exclusion – perhaps by starting the IPHC index in 2000. If the resulting index is found unlikely to be biased, then I recommend inclusion in the model with additional sd estimated.

Aleutian Islands

Most of the effort of the meeting was directed towards investigation of the properties of the EBS assessment model, as an SS assessment is already the agreed approach for that region. The AI is currently a Tier 5 that essentially applies a smoother through trawl survey estimates of total biomass. However, the assessed trend in biomass is less important than the most recent estimate in the provision of management advice. It was hoped that if reasonable approaches to data and modeling can be determined for the EBS, then many of those same approaches could also be applied to the AI region. My initial thought was that an agreed EBS model could be entirely transferred to the AI, but it was shown during the meeting that simplification of the AI model can lead to improved model behavior – particularly regarding retrospective patterns. Indeed, the removal of time-varying factors can sometimes improve retrospective behavior, possibly in conflict with general conclusions of recent publications (e.g. "when retrospective patterns are observed in a stock assessment, they are often corrected by introducing estimation of a time-varying parameter (usually selectivity, M or q)", Hurtado-Ferro et al. 2014).

The AI model is the same as for the EBS in that the trawl survey selectivity appears to be domed and to the left of the fishery, and that the IPHC survey has potential use for providing an index for older age-classes (Figure 3). Even without estimation of an additional sd, the IPHC index can be reasonably well fitted by the model, with 2012 being the largest influential residual. Further work on choice of a more appropriate selectivity function other than double-normal (or by changing the freedom of certain double-normal parameters) would probably improve the overall fit to IPHC lengths (Figure 4).

Before deciding to include the IPHC longline index and associated lengths in a proposed central AI SS model, an investigation into the properties of the AI IPHC 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.

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

The primary aim of the NMFS longline survey is to collect abundance, composition and bycatch information for Sablefish. Again, a desktop study has not been made to determine whether the survey is potentially biased with respect to Pacific cod abundance. Indices for Pacific cod are available for EBS and AI, although the survey does not cover the western AI region. Age compositions are not collected for Pacific cod by this survey, but there are many lengths collected.

During the review a question was raised about the possible over-weighting of surveys, particularly through the use of multiple longline surveys in a single model, and it was suggested that they could potentially be combined before addition to the model. My own preference on this is to keep independent data sources separate, and to let additional sd estimation weight each based on goodness of fit with all other data sources in the model. I think it is an advantage if independently collected indices show similar trends for the same size/age classes in the population, and should therefore receive more weight in those circumstances. Alternatively, conflicting indices should be down-weighted in an objective manner.

Bering Sea

Aggregated lengths for the EBS show that the NMFS longline survey seems to catch about the same size fish as the longline fishery, but not as many of the very largest fish as does either the fishery or IPHC surveys (Figure 1). Relative index trends show that the NMFS longline seems potentially more consistent with the shelf survey than the IPHC survey if shifted several years to the left (Figure 2). The NMFS longline survey does not show large changes in abundance that are biologically implausible as the IPHC survey does. Addition of the index to the model even without additional sd estimation shows a reasonable

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).

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

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.

Should the models be structured with respect to season?

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.

What constraints, if any, should be placed on survey selectivity at older ages?

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.

What constraints, if any, should be placed on survey catchability?

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.

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

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.

How should the various data sets be weighted?

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.

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

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.

Should the models be structured with respect to season?

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.

Should the models be structured with respect to gear type?

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.

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

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.

What constraints, if any, should be placed on survey selectivity at older ages?

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.

What constraints, if any, should be placed on survey catchability?

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.