SSC Review of the Northeast Region Discard Estimation Methods for Groundfish Quota Monitoring and Annual Catch Limits

Review provided by James Bence (review panel member)

General comments and summary

It was clear that substantial thought and care was put into analyses to determine an appropriate discard estimation methodology by Northeast Region and Northeast Science Center Staff. Many of my specific comments below focus on what more could be done to shore up limitations of the analyses to date, or how to better explain the content of the current working papers if they are recast in a technical report. These comments should not be interpreted to indicate an overall deficiency in the work, which I found truly impressive. I think the work presented in the working paper provides a strong foundation for using a cumulative annual estimator in the upcoming fishing season. In this regard I agree with the overall evaluation of the performance of this estimator as summarized in working papers 2 and 3.

Despite my favorable comments above, I do not think that the current suite of analyses should be viewed as establishing a design-based cumulative annual estimator as the method of choice for the long-term. The methodology used for discard estimation should be reconsidered once several years of data are available under the new sector management system. While it is reasonable and pragmatic to strive for a single design-based estimator in the short term, it may be the case that a specific different estimator could perform better for some specific sectors, and analysis of this possibility seems warranted for larger sectors once additional data are available. I am also concerned that even an annual cumulative estimator will run into problems when faced with some of the smaller sectors. As was discussed during the panel review meeting, alternative model-based approaches (also reference small-area estimation methods from survey literature) may perform better under situations of very limited data. I recognize that there are some policy issues here, in that a model-based method might attempt to improve estimates by "borrowing" information across sectors, and sector management is based on the presumption that what is done in one sector will not influence management in others. This noted, it may turn out that discard rates in some sectors will be too poorly estimated using only within sector and year data. This will become more evident as the new management system is put in place and may motivate some pragmatic model-based approaches in light of policy considerations. Additional future work (after future year's data become available) should also consider the influence of the transition methodology on the performance of estimators. Perhaps application of something like the adaptive estimator illustrated during the review panel and described in the supplemental report by Lanning could also solve problems associated with very small observer sample sizes. In this case it may be that the adaptive approach largely retains a prior estimate in cases where the cumulative estimate varies widely as a small sample size accumulates.

In general discards are a modest part of the total fishery kill and this likely explains why the timing of seasonal closures were fairly robust across discard estimation methods. Although there is no doubt this is generally true as summarized in working paper 3, it is also the case that discards are sometimes substantial. It may be prudent to consider specific higher levels of observing discards in cases where discards make up a higher proportion of the fishery catch, especially when the stock in question is at lower abundance that is of management concern.

Comments in light of each point of the review terms of reference

1. Compare and contrast the performance with alternative estimators of total discards with respect to precision and accuracy.

Here I consider the moving-average estimators versus the combined and separate estimators. I do not focus on the different temporal stratifications given that this is considered under #3 below. The simulations presented in working paper 2 clearly demonstrated that the moving average estimators experienced considerable bias (average percent discard consistently different than zero, usually too high). Furthermore the distributions of percent deviations from the truth were broader for the moving average-methods than for other methods except when observer coverage rates were far above what is feasible. Somewhat surprising to me, this bias was generally present even when the input discards were ordered randomly with respect to their original order. Possibly this is a small sample size effect, but whatever the cause it seemed clear to me that the moving average discard rates are not competitive with the other methods.

I agree with working paper 2 that for the simulations presented in that paper the cumulative to date estimator had the best performance overall. In a number of cases the distributions of errors (mid-year peak and end of the year estimates of discards) were quite similar among the methods. I agree that there was a tendency for the distributions to be tighter for the combined and stratified by quarter methods in the face of temporal trends in discards, but the most evident difference was a substantially tighter distribution of errors for the cumulative method for the Gulf of Maine Cod example when observer coverage rates were less than 40%, which is clearly a difference that would matter.

2. Evaluate impacts of trimming observations (e.g. large discard events) on the magnitude of bias and measures of precision.

Simulations presented in working paper 2 demonstrate that trimming produces substantially biases and such trimming should not be done. In a situation like this where large discard values and zero values have an important influence on the mean and the mean is the target of estimation it seems evident that trimming these values should not be done and the simulations support this view.

3. Examine the cumulative method and temporally stratified cumulative method (with various time steps and discarding patterns) and recommend a preferred method with consideration of the following:

   a. Within-year pattern of precision and bias, including a review of historical data used to estimate landings and discards (precision/accuracy)
   b. Feasibility/practical aspects, particularly implications of stratum size for small sectors and the ability to derive obtain fine-scale measures of total discard and its uncertainty.
   c. Estimate the probability of premature closure.

Among the cumulative methods, I prefer and recommend the annual cumulative method. Simulations conducted in Working Paper 3 show a general superiority of this method (under the conditions of the simulations in that paper) when considered over a suite of performance metrics and simulation scenarios.

Other stratifications did outperform the annual method for some metrics and under some specific conditions. For the low discard variation example of the GB longline haddock fishery, all the stratifications had similar spreads in the distributions of errors (total discards based on within 5% and

interquartile range metrics) and all had modest to small bias, with different methods winning for different methods by small margins across fleet sizes. For the high discard variation example of the GM gillnet haddock fishery the monthly stratification had the lowest spread in the distribution of errors for discards at low fleet sizes, the quarterly stratification had slightly narrow distribution for large fleet sizes. For the high discard variation example of the witch flounder trawl fishery the annual stratification retained the narrowest distribution of errors in annual discards. Other methods did win on bias but these differences were tiny and all the methods were essentially unbiased.

Generally the annual method consistently was outperformed on the metric of number of weeks that discard totals were adjusted downward. But in general the number of adjustments was not large and when these adjustments occurred they averaged less for the annual method. Generally speaking the concern that the annual cumulative method might lead to substantial within season adjustments of practical importance to the fishery (e.g., allowing harvest after when the fishery should have been closed or suggesting the fishery is near closure when it is not) does not seem to be a serious concern. The annual method generally was the winner across a range of conditions considered in the simulations for metrics that relate to practical use of the method. The discard rate itself was calculable more often as was its analytic variance estimate. It seems likely that this advantage of the annual method could become quite important for smaller sectors with small fleet sizes and relatively few observed trips on a monthly or even quarterly basis.

One metric reported as being to the advantage of the monthly and quarterly stratifications, namely the mean CV, is problematic. I have two issues here. First, CV was calculated as the variance summed over weeks with estimates over total estimated discards, but there are some weeks with discard estimates and no variance estimates. Including these weeks in the total discards will tend to inappropriately lower the CVs for the shorter stratifications. Second, I have some concern that the analytic variance estimate itself will be biased when sample sizes are smaller and this will be more common for the shorter stratifications. While obviously we would want to have discard estimates with lower CVs, I am afraid this is a case where estimation issues lead to potential and even probable difference between estimated and actual CVs.

A major limitation of the simulations reported in working paper 3 is that trips were selected at random for simulations, without respect to the time of year that they originally occurred. What this means is that the simulations demonstrate the general superiority of the annual stratification for a range of sample sizes, variability in discard, and other underlying characteristics, in the absence of seasonal patterns or autocorrelation in discard rates. However, it is under conditions when different temporal strata have different discard rates that one would expect temporal stratification to provide advantages. I still ended up recommending the annual method because working paper 2 suggested that even quite strong trends would at best produce modest advantages for quarterly versus annual stratification, combined with the pragmatic fact that monthly and quarterly stratification would be problematic due to limited sample sizes in some cases. It would be worthwhile to follow up the simulations of working paper 3 with some checks on robustness in the face of seasonal patterns. E.g., how would the annual versus quarterly versus monthly stratifications perform if randomly selected discard values were multiplied by monthly adjustments to impose an underlying “true” seasonal pattern to discards?

<u>4. Provide guidance of methods to measure uncertainty of the preferred method (e.g., asymptotic vs. bootstrap estimates of variance).</u>

Working paper 3 provides a comparison of analytical and bootstrap estimates of CVs. These results show that for the conditions of these simulations that the two estimates are highly correlated. It appears that in all the simulations the bootstrap estimate of the CV was higher than the analytic estimate and that the difference became larger as the analytic estimate increased. It may be the case (not possible to be sure from the plot) that the two estimates are close to proportional.
The working paper tends to talk about the higher value for the bootstrap estimate as reflecting bias. Here I think it is important to keep in mind that both the bootstrap and analytic variance (and hence CV) estimates are subject to potential bias at low or moderate sample sizes. Thus comparing the two does not provide much information useful for picking between the two.

One possibility would be to work toward estimating the ratio between either of these estimates of CV and the true CV. Possibly one or the other is close to unbiased and could be chosen on that basis. In theory the true CV could be estimated from simulations like those that were done. Variance is defined by E([True-expected]^2). For each simulation both true and expected are obtained. From a collection of such results one could estimate the variance. The difficulty in obtaining a estimate of variance from the simulations that were done is that "true" varied among the simulations, and its possible that the variance or CV or both varied as the truth varied. True varied because the actual discard sequence was randomly selected for each simulation.

Perhaps the existing simulations could be used to approximate the true CV. One approach would be to calculate "true" variance as the average squared error over simulations and dividing the square-root of this by the average true value. Alternatively one could calculate relative errors ([estimate-truth]/truth) and use the standard deviation of this as an estimate of the true CV over the simulations. Both approaches presume that variation in the truth among simulations are not so large as to dramatically influence the true CV. Comparison of average bootstrap and analytic estimates of CVs with such an estimate of true CV could provide information on the proportional bias of each. Another option would be to do additional simulations where for each of a suite of different randomly selected true discard sequences, one chooses repeat observed samples so as to calculate the true variance of the estimation method for each "true" discard sequence.

There are of course other estimators than the analytical and bootstrap estimators that were reported on. Most notably, there are other variants of the bootstrap (bias corrected, accelerated, various adjustments for finite samples). Ultimately, I was left with the question of how the CVs would be used with regard to actual within season management of the fishery. I have concerns that the bootstrap (any variant) is still subject to bias and may be operationally problematic, given the amount of computing required. Given that the bootstrap and analytic estimates were highly correlated and I do not know which is better it seems best to me to use the analytic method, possibly with an adjustment for a assumed known proportional bias (perhaps estimated as described above). I put forward this suggestion with some hesitation because ideally you would want to examine by simulation which of the two (or alternative) estimators tracks the truth best, and this would include any ad hoc approach to adjusting the analytic estimator.

My understanding of the potential within season use of the CV estimates is that they provide a flag to managers regarding interpretation of the discard total estimate. E.g., a rapid jump in the discard estimate combined with a high variance might suggest the estimate is not well established. It might even lead to an increased sampling of discards. I would think this kind of use of the CVs will generally be a response to the relative values, and will not be substantively influenced by modest bias. If this is the usage, I would think that an end of year refined estimate of CV (say complicated bootstrapping

approach) should be possible even though within season analytic (possibly adjusted by multiplier) estimates were used for pragmatic reasons. It might be the case that the absolute variance or CV is more important at that point if this variance feeds into a stock assessment procedure.

5. Provide guidance on risks of alternative management actions given the uncertainty and/or behavior of the preferred discard estimation method. In particular, consider the costs to industry/sector from premature cessation of fishing and the risk to the resource from excess harvests.

Discard estimates appear to be subject to quite large errors relative to the magnitude of the discards (i.e., CVs are large). However, the information provided to the review panel suggests that the proportion discards make up of the total fishery kill is likely to be small in most cases. In such cases the uncertainty in discards has relatively little influence on when fisheries will be closed and there is little chance that poor estimation of discards will lead to undue fishing mortality so as to put stocks at risk. This might be better communicated by also reporting the CV of total fishery kill (assuming known harvest), which would illustrate that the total fishery kill is generally well known (to the extent harvest is) even in the face of the uncertain discards. It does appear that when stocks are subdivided into multiple sectors, with separate seasonal management, that there is more risk of early or late closure for individual sectors than there is for overall overfishing. This is because the sector specific errors would tend to cancel out in their effect on the stock. This said, it is still the case that with discard rates like those used in the simulations presented to us these closure issues seem modest. Those sectors expected to have substantially higher discard rates may need special attention (higher sampling coverage, more precaution especially early in the year).

6. Consider implications of finer-scale stratification on performance of estimators that might be required for sector-specific discard rates and will be required for multi-stock species.

Finer scale stratification leads to fewer vessels in a stratum and thus at any given observer coverage smaller numbers of trips observed for discards. Here my assumption is that in general finer-scale stratification in this part of the terms-of-reference implies finer-scale definition of sectors. To the extent that separate trajectories over time of discards are required for a stratum for management purposes (my assumption regarding this item in the terms of reference), the simulations demonstrate that the lower sample sizes associated with smaller fleet sizes will generally tend to favor the annual temporal stratification estimator.

The simulations did not directly address the situation where analysis stratification is finer than sectors used for management. For example one might define several mesh size strata within a sector, but then the estimates would be combined (as in equations 1&2 of working paper 3). From general principles I would think any such additional stratification in other dimensions would tend to favor improved relative performance of estimators using larger temporal stratification, including pragmatic issues such as simply being able to calculate the needed ratios for each stratum. It seems clear that if differences in discard rates tend to be larger among such mesh size strata (or other subcategories such as vessel type) than among monthly or quarterly temporal strata then use of such stratification would be preferred over temporal stratification.

Other Comments

1. Clarifications about the estimators. As was discussed during the panel meeting, the exact definitions of the different estimators of discards were not entirely clear to the panel, based on the presentations in

the working papers. Although the discussion of this issue during the review panel meeting was useful, what I thought was clear after the panel meeting was no longer 100% clear as I revisited the written reports. My major concern had to do with the combined ratio estimator. I had two issues here.

First the definition of the estimator (the second Eqn. 2 on page 3 for the combined estimator of working paper 2) suggests that the numerator and denominator terms are summed over all strata. Strata that are actually combined over are not clearly defined in working papers 2 or 3. Working paper 1 refers to stratifications such as sector, gear, mesh and stock area in direct reference to the application of discard methodology described in working paper 2 (see page 7, scenario 4). I actually think in working papers 2 the combined estimator is not applied across these stratification variables but I am not certain. The examples for which results are presented already seem to be focused on a particular segment of the fishery for a stock. Discussion during the working meeting suggested that there was probably no combination over sectors.

Second, the temporal strata used for the combined estimator in working paper 2 were not (as far as I could tell) explicitly defined. I heard and verified that the combined and separate ratio estimators would produce identical results when there was a single stratum the two were calculated for. Thus it seemed evident to me, given the different results for the combined and annual separate estimators in working paper 2, that either the combined or separate estimator were being combined over more than one stratum. This is basically why I concluded that in working paper 2 the combined ratio estimator was calculated on the basis of quarterly strata, to obtain a single ratio used to obtain a discard estimate for the specified data. I am, however, uncertain about this. I apologize if I missed when this was clearly explained during the panel meeting, but the need to be very clear and consistent about the estimators in future reports is perhaps highlighted by my comments here.

2. Generally, it would have been easier to read and integrate the working papers if the motivation for the different estimators that were considered were clearly laid out (why was a estimator a contender and considered in simulations?). There is some of this (e.g., because it was used in stock assessments…) but more would have been helpful. I found this especially lacking for the combined ratio estimator. This essentially uses a ratio estimate to obtain discards in a stratum and kept in a stratum (number of trips x discards per trip, number of trips x kept per trip) and sums the totals of these estimates to get the numerator and denominator of the overall ratio estimate. Frankly I don't see much appeal to this estimator. It seems as though one needs observations of discards in a stratum to include it in the calculations (without samples one cannot calculate discards per trip) negating one of the apparent advantages of the estimator. I would think this estimator poses difficulties in variance calculations. It also seems likely to produce at least "by stratum" biases if there are differences among strata in discard rates (and if there were not why stratify at all?). Perhaps there are potential advantages reported in the literature and summarized in the cited report but I think a a few sentences summarizing these would have been helpful. I think this could be important for any eventual technical report.

3. Overall I found a number of the choices that were made when going from working paper 2 to working paper 3 puzzling although most of this was cleared up during the review panel meeting. Any ultimate summarization of these results in a technical report will need to carefully lay out how the results from the simulations in working paper 2 motivated working paper 3. Why was the combined estimator no longer considered? Why was the focus on performance in the face of temporal patterns in discard patterns from Working Paper 2 not followed up in Working Paper 3?

4. There were a few places in text and interpretation in working paper 3 where it would have been nice to be more explicit that output from equation 3 is a variance estimate not the variance. E.g., I was uncomfortable with CV being used as a metric of performance across different stratifications given that the estimator itself is known to have sample size related biases. The metrics that directly evaluated how close point estimates (that would be used for management purposes) were to the truth or how variable the actual errors in the point estimates were among simulations seemed more useful.

5. The basic difference between bootstrap and analytic CVs was a bit puzzling to me. It appears that the bootstrap estimate of variance must always have been larger than the analytic estimate. The explanation of this on page 9 of the working paper 3 was not convincing to me. It could be worthwhile to seek a better explanation for this effect as it may help in choosing between estimators.

6. Working paper 3 provided fairly compelling evidence that 1000 simulations were sufficient to evaluate the estimators, and that results were robust to frequency of discard calculations (weekly or daily).

7. As discussed during the panel review it would be worth verifying that calculations based on species hailed weights on amounts kept for observed trips provide similar answers to those using dealer weights (e.g., as in working paper 1, equation 6). On reflection it is not actually critical if there is a consistent difference between hailed and dealer weights provided that any such influence on observed kept values also applies to observed discard values.

8. I am not sure I agree that it is a policy decision to decide whether one needs more than one observed haul from a subtrip in order to consider it observed (see text near top of page 10 of working paper 1). It seems to me that this is a topic subject to analysis (e.g., what would happen in simulations if you treated a subtrip as unobserved if only one haul was observed?). I would think any such definition should depend in part on what fraction of the total number of hauls was observed not just the number.

9. I had some difficulty following the CV estimation methodology of working paper 4 and clarity would have been helped by inclusion of explicit equations. I was not very concerned by this because the moving average estimators were not viable contenders.

10. Working paper 5 provides a clear description of the transition discard methodology. This describes how the discard rate assumed at the beginning of the year will be updated as in-season data become available. It seems to me that there could be some interaction between the rate at which this transition occurs and the performance of different estimators. However I would think in general simulations that imposed a slower transition rate would tend to favor annual stratification. As long as the transition is essentially complete well before the end of a temporal stratum then the ultimate performance of the method in terms of quality of discard estimates will not be impacted and this condition is more likely with longer temporal strata. It is worth noting that the transition rate methodology has some potential to influence the within season estimates of total discards, which were represented in several of the performance metrics.

11. As discussed during the panel review, the choice of an appropriate alpha for the transition methodology should not be based solely on how many samples are required for desirable estimation properties for the cumulative ratio estimate. The concern here is that the method in essence averages an initial discard rate with the cumulative estimator. It is possible that the initial rate was based on a modest sample size (e.g., 12). Perhaps a rule for choosing a fixed alpha should account for how many

data were used in the initial discard estimate. I would think that alpha should be such that the initial estimate was weighted less than or equal to the in season cumulative estimate when the in season sample size equaled the sample size used to calculate the initial rate.

12. I read over the additional working paper describing the adaptive alpha method. While I found this method intriguing, I would not recommend its use at this time until additional testing of its performance is possible over a wide range of conditions. I do not provide a full review of the paper describing this method because I was not able to fully understand the details of how the method was applied. The overall clarity of this paper was remarkable given the short time period available to prepare it, however the definitions of some of the variables were not very clear to me (e.g., I was not entirely sure whether A an F referred to actual and smoothed forcasts of the cumulative discard rate). I was not particularly concerned that alpha does not go to zero for large sample size as alpha raised to the power of sample size does go to zero. It may be practical to simply decide on a lower limit to alpha and an upper limit after which the transition rate is replaced by a simple cumulative estimator.

13. There was some discussion during the panel meeting regarding the assumption that all discarded fish die. It seems a simple matter to replace this assumption with some other assumption about survival when calculating total fishery kill ((1-survival rate)discards + harvest) when such information is available for a sector, although this could grow complex if the survival rate is assumed to vary with attributes such as fish size or age or mesh size. It will be important to carefully monitor amounts of discards, and amounts of discards assumed to die (influencing in season management). One can envision situations where the current stock assessment, using the latest available information would make different assumptions regarding the survival of discards than were made during each previous year for in-season management, which may have changed over time. This seems reasonable, if the difference is based on availability of new information, but should not reflect inconsistent and unjustified analyst choice. In general the same assumptions should be made during a fishing season as are made in the subsequent assessment, and when changes are made in treatment of discards (either for an assessment or in-season management), due to new information, these changes should be translated into both these uses of the discard data.