Model averaging by cross-conditional decision analysis

NOAA FISHERIES
Alaska Fisheries Science Center

Grant Thompson

September 17, 2019

Background

Motivation for developing the approach

- A pair of Team and SSC requests:
  - BPT8: “For next year’s assessment, the Team recommended that ... the author considers bringing forward an ensemble of models to capture structural uncertainty with a justifiable weighting....”
  - SSC8: “...Moving forward, weighting of models for an ensemble may be developed based on the relative plausibility of each model hypothesis. The SSC recommends further efforts in developing this approach.”
- The most straightforward way to interpret the above requests is that the Team wants weighting to be objective and the SSC wants weighting to be subjective.
  - Objective: weights are computed statistically.
  - Subjective: weights are assigned based on relative believability of model structure or results.

Problem #1: ensembles require weighting

- Basic steps in a Bayesian approach:
  - Choose a quantity of interest.
  - Calculate the posterior distribution of the quantity of interest.
  - Choose a loss function.
  - Integrate the product of the posterior distribution and loss function.
  - Minimize the integral (i.e., the risk).
- Step 2 in the above list becomes complicated when dealing with an ensemble (i.e., a set of models), because the posterior distribution will be an average of the *model-specific* posterior distributions, but there is yet no consensus on how this average should be computed
  - Equally or unequally weighted?
  - If the latter, how to specify the weights?

Previous solutions to problem #1 (1 of 2)

- Many authors suggest that the weights should ideally consist of Bayesian posterior probabilities.
- However, computation of such probabilities can be difficult and, more importantly, requires that the same data be used to fit all of the models in the ensemble.
- Although some studies have successfully produced fully Bayesian probabilities for the models in an ensemble, most have defaulted to approximations (e.g., purely subjective “plausibility weighting” or weights based on importance sampling, Akaike Information Criterion, Bayesian Information Criterion, Deviance Information Criterion, bootstrapping, cross-validation, or retrospective analysis) or assumed equal weighting.

Previous solutions to problem #1 (2 of 2)

- The “superensemble” approach, introduced originally by Krishnamurti et al. (1999) in the fields of weather and climate forecasting and recently applied to fisheries management by Anderson et al. (2017) and Rosenberg et al. (2018), provides another alternative, in which weights are estimated statistically so as to minimize an objective function, which, in a Bayesian decision analysis, would be the expected loss, although non-Bayesian objective functions could also be used.
- These two major alternatives, weights that *reflect probability* and weights that *maximize performance*, are not mutually exclusive.
- In fact, both can be used simultaneously, as they serve different purposes.
- The former are necessary to *compute* the expected loss, whereas the latter can be used to *minimize* the expected loss.

Problem #2: unobservable quantity of interest

- However, when ofl is the primary quantity of interest and an ensemble is involved, the methods that have been used for optimizing performance-based weights in other disciplines are typically not applicable.
- This is because, in other disciplines such as weather and climate forecasting, a time series of true values for the primary quantity of interest exists (e.g., precipitation is routinely measured with negligible error) and can be used to estimate ("train") the optimal weights, but in fishery management, no time series of "true" ofl values exists.
- One possibility is to optimize the weights by training on data that *are* observed, such as a survey index time series (as suggested by Stewart and Martell 2015), but there is no guarantee that an ensemble tuned to fit something other than the quantity of interest will be good at estimating the quantity of interest.

An approach that addresses both problems

- Major steps in the method developed here:
  - Treat each model in the ensemble, one at a time, *as if* it were true.
  - Compute the risk resulting from a performance-weighted average of the models in the ensemble relative to the best point estimate from the conditionally true model (the “pivot” model).
  - Multiply the results by the probability that the pivot model is “true.”
  - Sum across pivot models to obtain the risk for the entire ensemble.
  - Tune the weights so as to minimize the ensemble risk.
  - Use those weights to create an ensemble distribution from which the optimal value of the quantity of interest can be estimated.
- The above process provides a *cross-conditional decision analysis*.
- Some similarity to the superensemble approach, in that a statistically tuned set of weights is computed, but goes beyond that approach to account for the fact that no time series of “true” values exists.

Probability density (mass) functions

- In the approach developed here, uncertainty is represented in the form of a probability mass function (pmf)
- Steps in generating the pmfs:
  - Fit each of the *nmod* pivot models to the "real" data.
  - Generate *nsim* sets of bootstrap data from each fitted pivot model.
  - Fit each candidate model to each of the $nmod \times nsim$ bootstrap sets.
  - Procedure results in a total of $nmod \times nsim \times nmod$ model runs.

Loss function

- The following loss function is assumed:

$$loss(y|\hat{y}, ra) = \left(\frac{y^{1-ra} - \hat{y}^{1-ra}}{1-ra}\right)^2$$, where:

  - y is the quantity of interest,
  - $\hat{y}$ is intended to approximate the true-but-unknown value of y, and
  - ra is the level of risk aversion, where:
    - any value of $ra > 0$ implies true risk aversion
    - the special case of $ra = 0$ implies risk neutrality, and
    - any value of $ra < 0$ implies risk proclivity.
      - Here, risk aversion means that any underestimate is preferred to an overestimate of the same magnitude.

Risk (expected loss)

- The procedure is fairly general, and should be applicable to a wide range of choices as to the quantity of interest, with two constraints:
  - the quantity of interest cannot take negative values, and
  - if any value of ra other than 0 is chosen, the scaling of the quantity has to be consistent with the meaning of risk aversion.
- Risk is defined as the expected loss (i.e., the sum of the product of the pmf and the loss function).
- The risk-minimizing value of $\hat{y}$ is the y mean of order $1-ra$, defined as the $(1-ra)$th root of the $(1-ra)$th noncentral moment of the y pdf.

$$m_y(1-ra) = \left(\int_0^\infty g_y(y) y^{1-ra} dy \right)^{1/(1-ra)} .$$

- If $ra=0$, solution is arithmetic mean; if $ra=2$, solution is harmonic mean

Example application

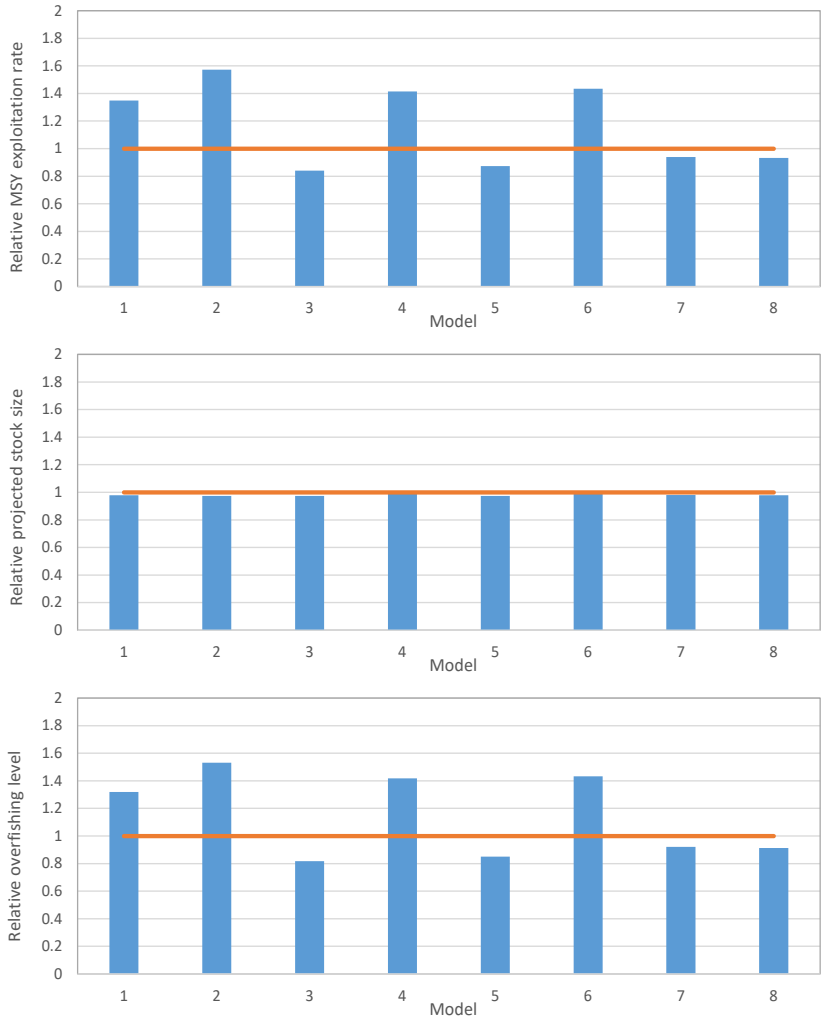
Description of example (1 of 2)

- As a test case, CCDA was applied to an ensemble of simple surplus production models, with ofl chosen as the quantity of interest.
- The most important feature of the ensemble is that the natural mortality rate is, potentially, a function of up to nv environmental covariates, where the values of all values in the time series of all environmental covariates were assumed to be measured without error.

Description of example (2 of 2)

- An ensemble of eight models was created by setting $nv=3$ and using a full factorial design as follows:
  - Model 1 included no environmental covariates.
  - Models 2-4 included exactly one environmental covariate:
    - Model 2 included covariate #1.
    - Model 3 included covariate #2.
    - Model 4 included covariate #3.
  - Models 5-7 included exactly two environmental covariates:
    - Model 5 included covariates #1 and #2 (chosen as “true”).
    - Model 6 included covariates #1 and #3.
    - Model 7 included covariates #2 and #3.
  - Model 8 included all three environmental covariates.

Results of base runs (*umsy*, *xpro*, *ofl*)



Some bootstrap results

- Out of 4000 bootstrap data sets ($nmod = 8$ models $\times$ $nsim = 500$ bootstrap data sets per model), all models were determined to have converged in 2594 instances (64.8% of all runs).
- The risk-minimizing value of ofl for a given model when fit to the bootstrap data sets generated from its *own base run* for each model (i.e., no cross-conditioning yet) is shown below, for both the risk-neutral (ra=0) and risk-averse (ra=2) cases, along with the estimate of ofl from the respective base run:

Model:	1	2	3	4	5	6	7	8
Base:	0.275	0.320	0.171	0.296	0.177	0.299	0.192	0.191
ra=0:	0.301	0.381	0.199	0.365	0.222	0.376	0.271	0.224
ra=2:	0.261	0.340	0.193	0.342	0.219	0.347	0.241	0.219

Optimal model weights

- The optimal model weights for $ra=0$ were:

Model:	1	2	3	4	5	6	7	8
w:	0.198	0.160	0.000	0.114	0.427	0.000	0.093	0.007

- The optimal model weights for $ra=2$ were:

Model:	1	2	3	4	5	6	7	8
w:	0.026	0.000	0.162	0.008	0.419	0.369	0.000	0.017

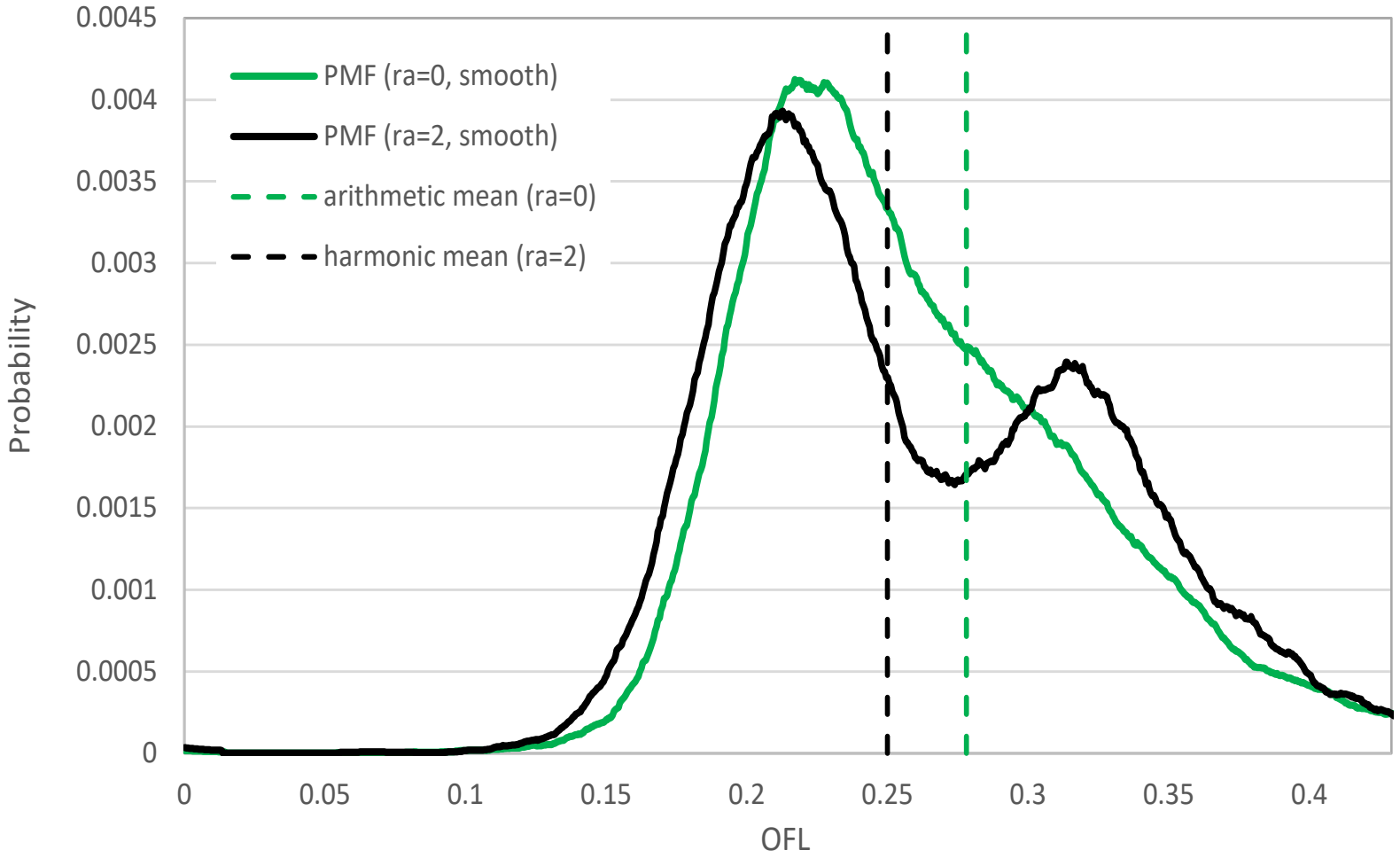
- Under both values of ra, Model 5, which was the true model, was given the most weight.
  - No guarantee that this will always happen!

Statistics of the optimized *ofl* pmfs

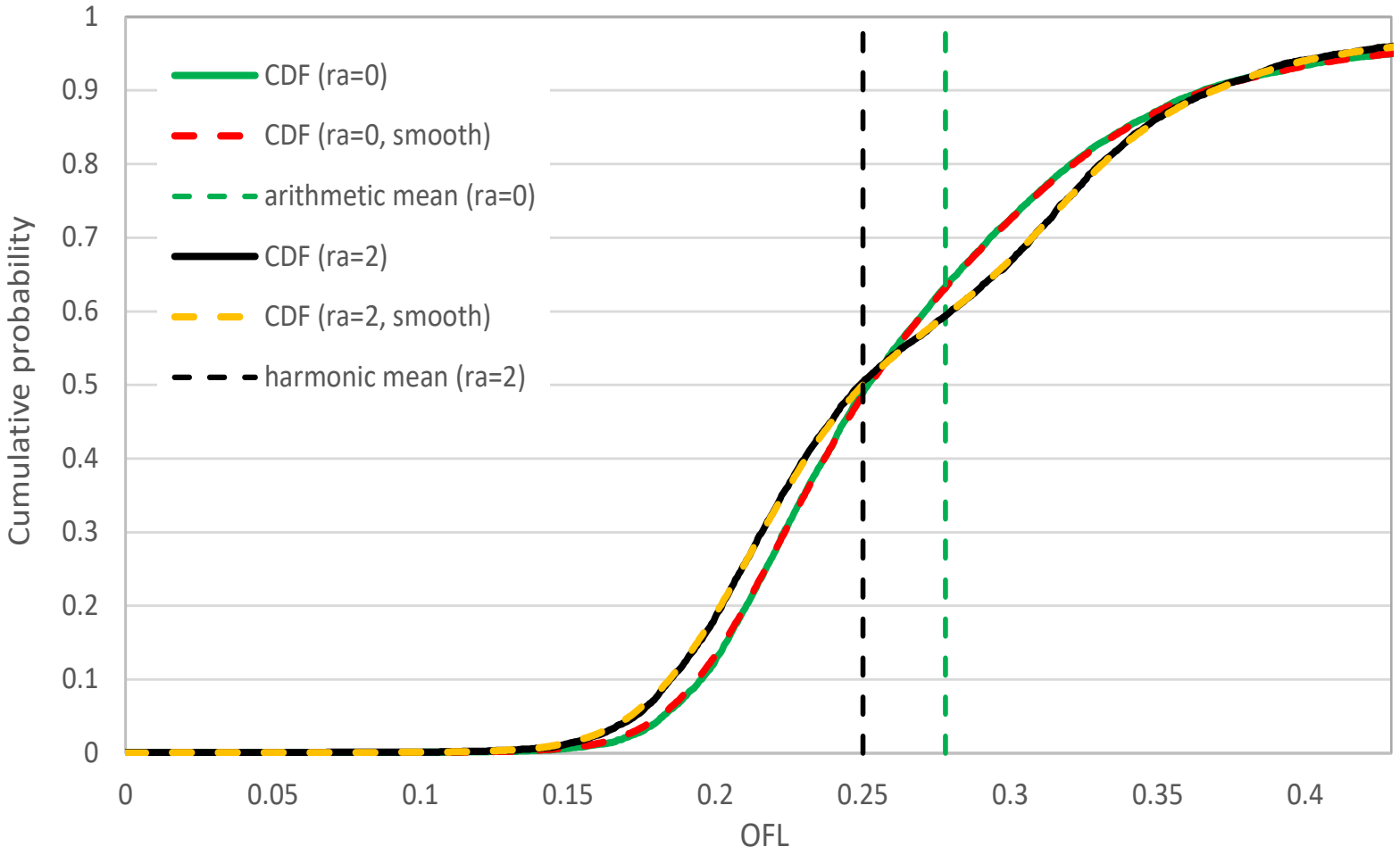
Statistic	*ra* =0	*ra* =2
median	0.252	0.249
arithmetic mean	0.278	0.275
geometric mean	0.264	0.261
harmonic mean	0.254	0.250
standard deviation	0.113	0.110
coefficient of variation	0.405	0.401
skewness	3.933	3.701
*p**	0.634	0.505

- It may also be of interest to note the cumulative probability associated with the risk-*averse* optimum as computed from the risk-*neutral* distribution, which is 0.490.

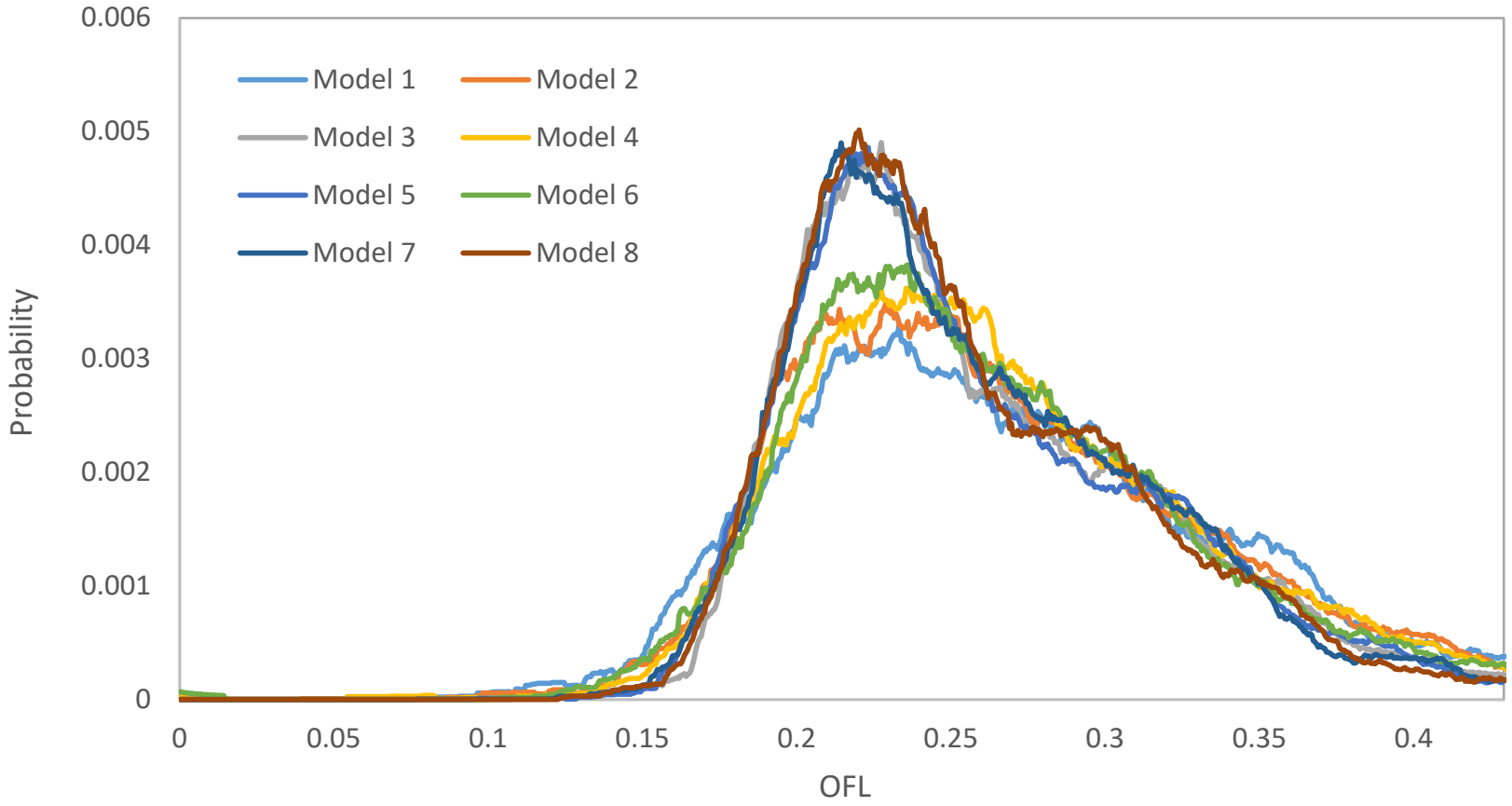
Optimal estimates of the *ofl* pmfs



Optimal estimates of the *ofl* cdfs

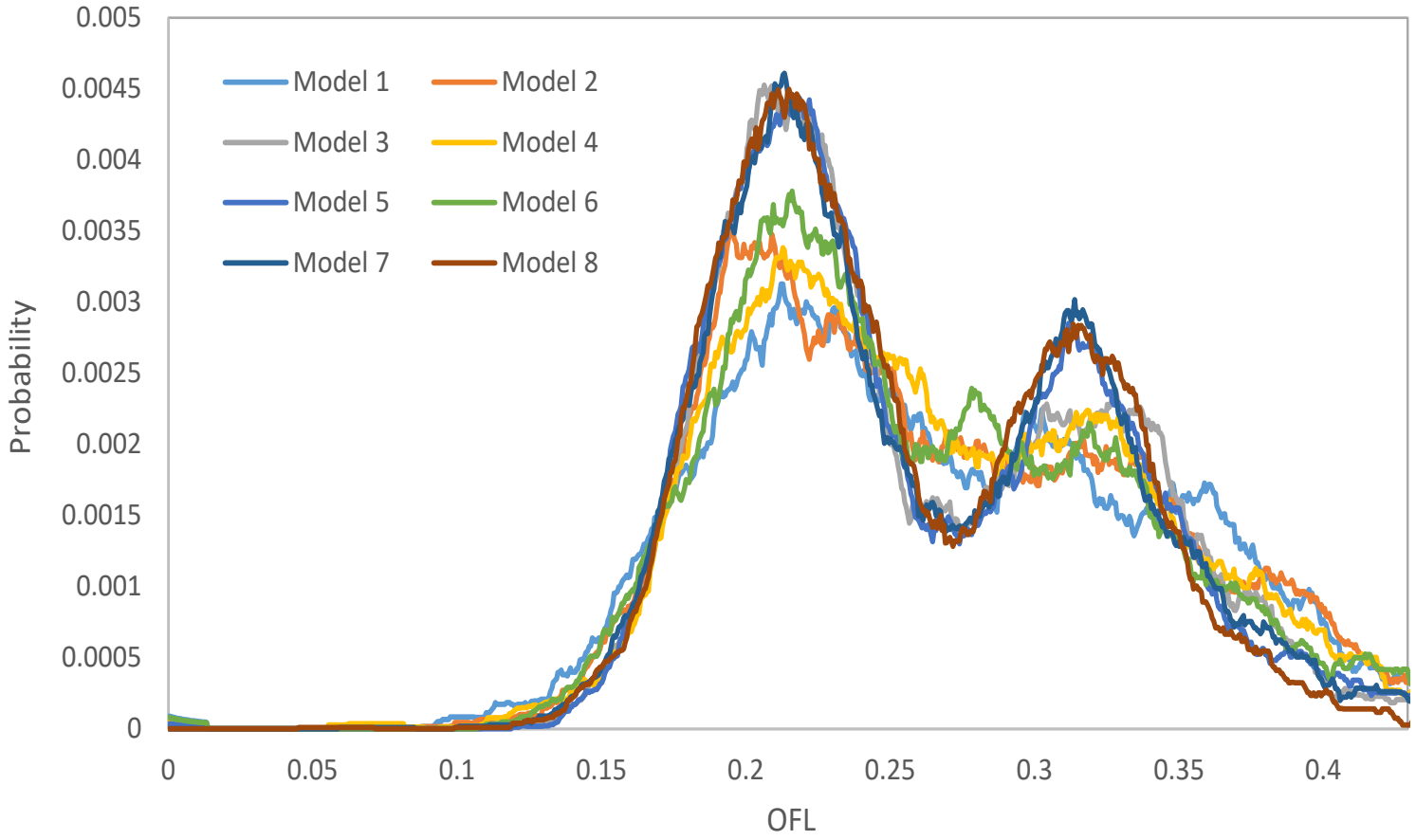


Unweighted individual *ofl* pmfs: risk-neutral

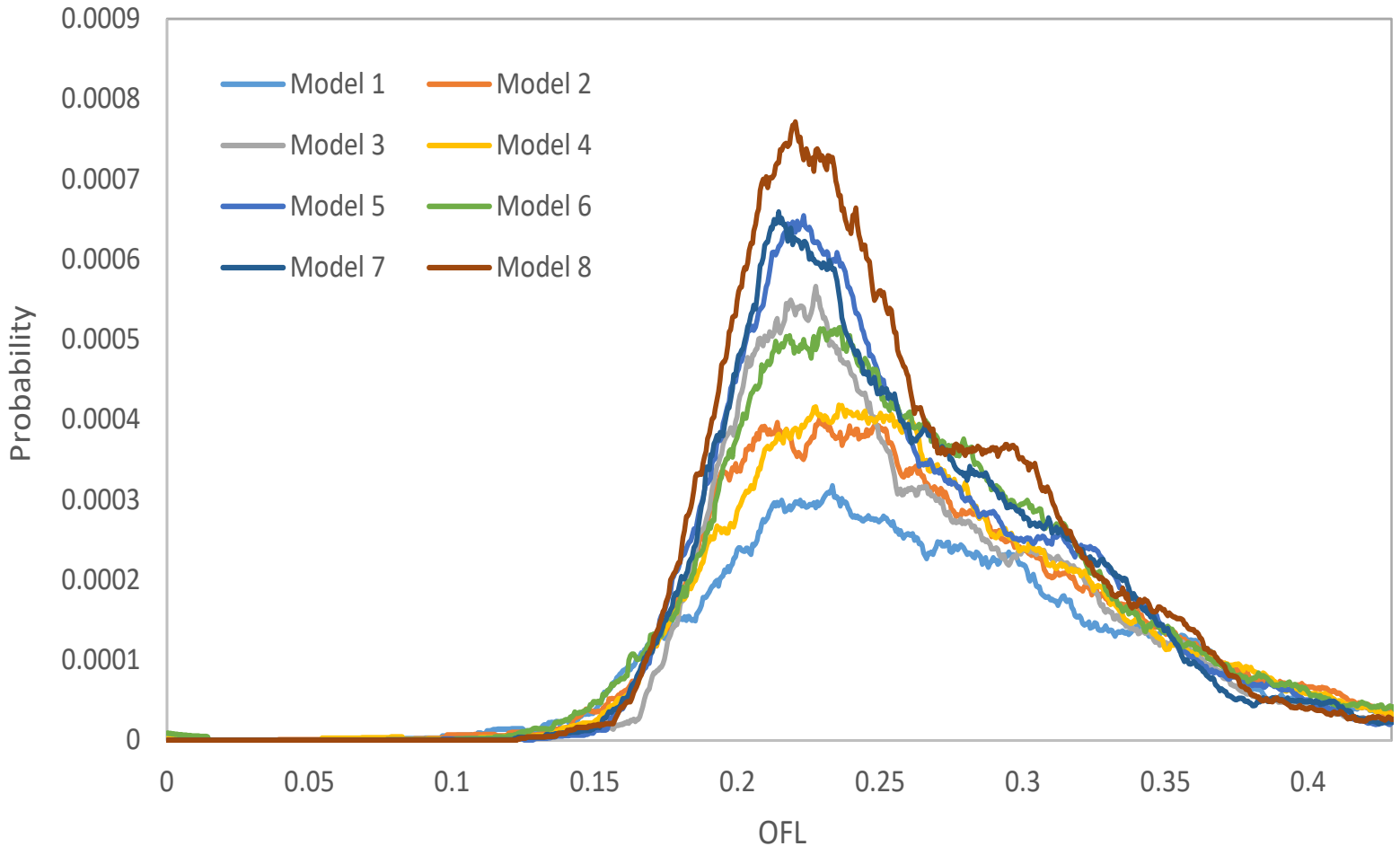


Unweighted individual *ofl* pmfs: risk-averse

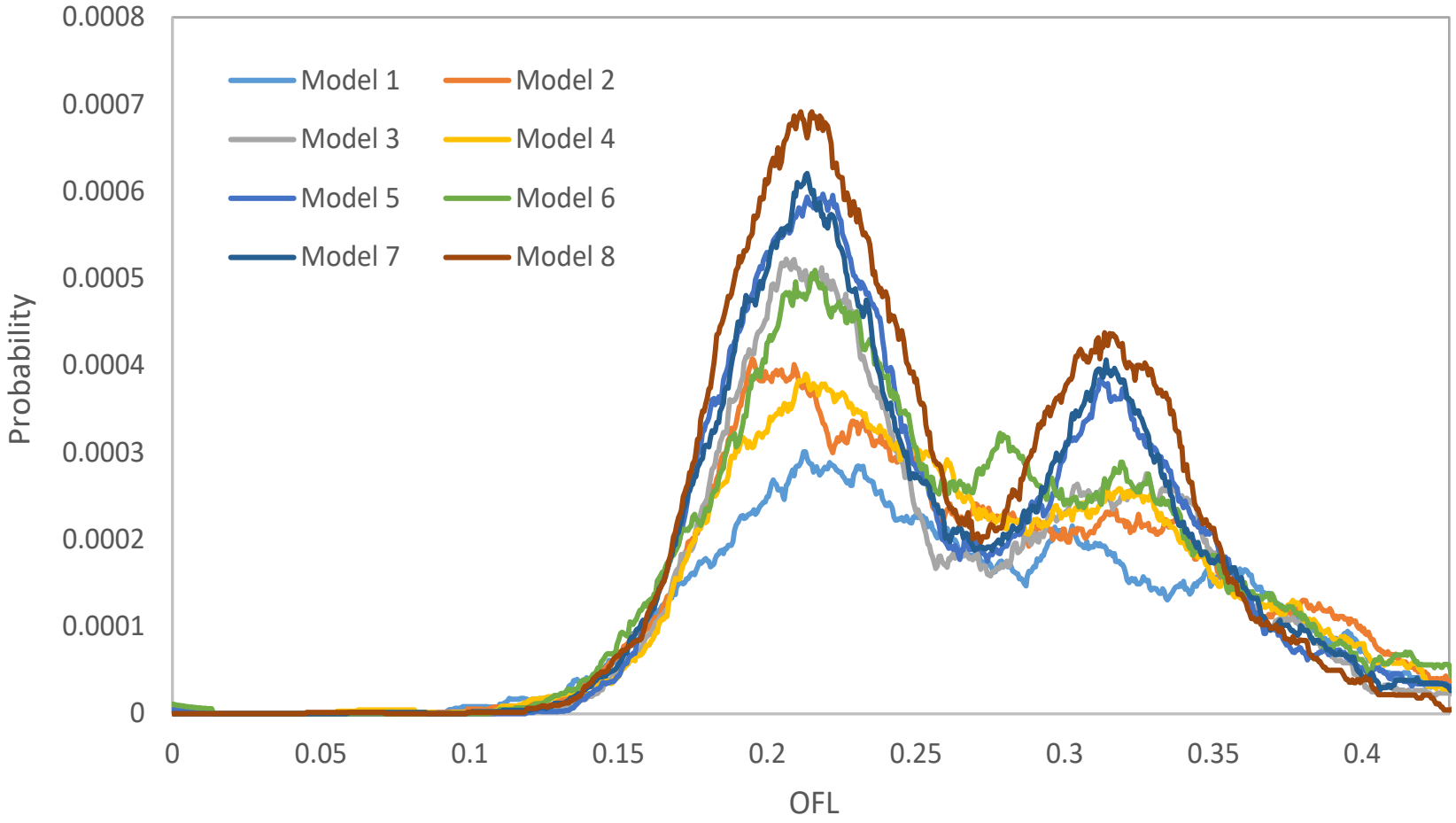
- Models 5, 7, and 8 contribute most to the bimodality.



Weighted individual *ofl* pmfs: risk-neutral



Weighted individual *ofl* pmfs: risk-averse



Cross-validation results (1 of 2)

- Mean and standard deviation of model weights from 10-fold cross-validation (repeated 10 times, with the folds chosen randomly each time).
- Mean weights from the training data sets are very close to the weights computed from the overall data set in both the risk-neutral and risk-averse cases.

Model	*ra* = 0			*ra* = 2		
	All data	Training data		All data	Training data	
		Mean	Sdev		Mean	Sdev
1	0.1975	0.1968	0.0228	0.0259	0.0273	0.0122
2	0.1603	0.1603	0.0126	0.0000	0.0000	0.0000
3	0.0000	0.0008	0.0063	0.1620	0.1583	0.0273
4	0.1142	0.1144	0.0103	0.0076	0.0100	0.0115
5	0.4274	0.4230	0.0189	0.4185	0.4191	0.0388
6	0.0000	0.0005	0.0027	0.3685	0.3637	0.0154
7	0.0934	0.0920	0.0088	0.0000	0.0000	0.0000
8	0.0073	0.0121	0.0124	0.0174	0.0216	0.0140

Cross-validation results (2 of 2)

- Mean and standard deviation of risk (expected loss) from 10-fold cross-validation (repeated 10 times, with the folds chosen randomly each time).
- As expected, the mean expected loss from the training data sets is close to the expected loss from the overall data set, while the mean expected loss from the testing data sets is slightly higher, and the standard deviation of expected loss is greater for the testing data sets than for the training data sets.

$ra = 0$				$ra = 2$			
All data	Cross validation data			All data	Cross validation data		
	Subset	Mean	Sdev		Subset	Mean	Sdev
0.0084	train	0.0084	0.0002	0.9447	train	0.9439	0.0120
	test	0.0085	0.0017		test	0.9631	0.1247

Discussion

Using CCDA to produce harvest specs (1 of 2)

- To produce an optimal estimate of the ofl corresponding to a particular level of risk aversion, the approach developed here involves three distinct levels of optimization:
  - Optimize the conditional (on each pivot model) ofl
  - Optimize the ensemble pmf
  - Optimize the ensemble ofl
- In the example presented here, results for two levels of ra were provided; the first corresponding to ra=0, representing a risk-neutral perspective and yielding an ensemble ofl of 0.278, and the second corresponding to ra=2, representing a risk-averse perspective and yielding an ensemble ofl of 0.250.
  - Approximate 10% buffer.

Using CCDA to produce harvest specs (2 of 2)

- The latter would a natural choice for the abc, defined in Federal guidelines as an annual catch based on a control rule "that accounts for the scientific uncertainty in the estimate of *ofl*, any other scientific uncertainty, and the Council's risk policy" (§600.305(f)(1)(ii)).
- Here, the "risk policy" would consist of a specified value of $ra>0$.
- Note that ofl is still the quantity of interest in the procedure used to produce an abc value; the difference is simply the level of risk aversion.
- Unless the FMP is amended, the estimate of abc resulting from this procedure would be subject to the existing $maxABC$ constraint.
- This means that the procedure would have to be repeated, with $maxABC$ as the quantity of interest, to determine if the risk-averse estimate of ofl is no greater than the risk-neutral estimate of $maxABC$.

Some issues with the approach

- Different from the common p^* approach
- Nothing like this is done for any NPFMC stocks currently
- Requires specifying each model's probability of being "true"
  - Compare to SSC request for weighting based on "relative plausibility"
- Requires specifying a level of risk aversion (for *abc*)
  - Compare to need for specifying p^*
- Complicated!
- Time-consuming ($nmod \times nsim \times nmod$ runs required)!
- Very small amount of testing to date
- Dirichlet-multinomial distribution may not yet be implemented in the SS routine for generating bootstrap data sets

Contrast with Team/SSC approaches (1 of 7)

- This approach differs significantly from the approaches recommended by the Team and SSC:
  - BPT8: "...All model outputs in the ensemble that are management related should be averaged, and the ABC should be determined from those averaged outputs (i.e., the application of the control rule to averaged biological reference values)."
  - SSC2: "...The combining of model output should occur on the basic estimates from the assessment (biomass, F, etc.) and not the reference points themselves."
- The steps involved in implementing the Team and SSC approaches are listed on the next 2 slides.

Contrast with Team/SSC approaches (2 of 7)

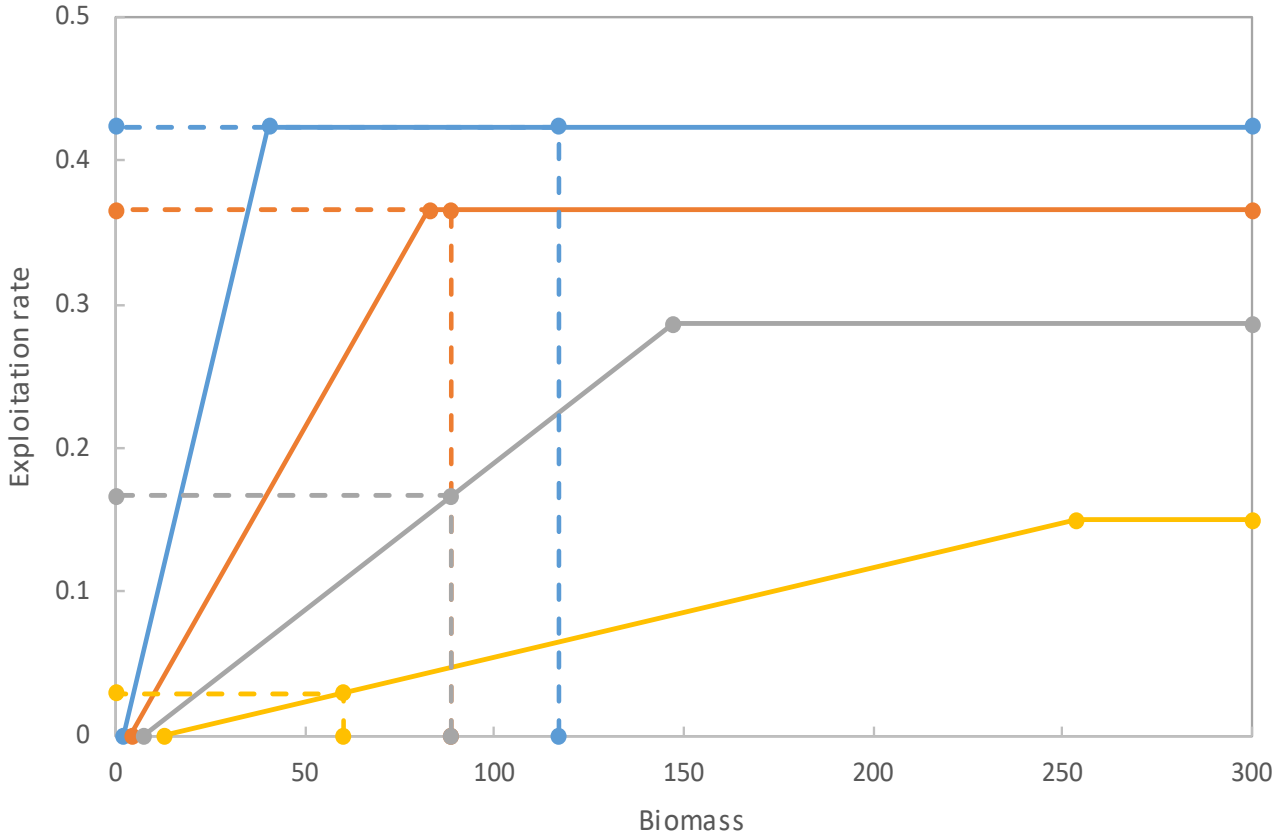
- Team's approach:
  - Compute averages of model-specific natural mortality rates, maturity-at-age vectors, selectivity-at-age vectors, and weight-at-age vectors.
  - Compute averages of model-specific $F_{40\%}$ and $B_{40\%}$ estimates.
  - Use the averages computed in step 2 to parameterize an "average" $maxABC$ harvest control rule.
  - Compute average of model-specific projected spawning biomasses; then insert that average into the "average" $maxABC$ harvest control rule constructed in step 3 to obtain an "average" $maxFABC$.
  - Use averages computed in step 1 and "average" $maxFABC$ value obtained in step 4 to compute an "average" $maxABC$ at each age.
  - Compute the "average" $maxABC$ as the sum (across ages) of the model-specific "averages" computed in step 5.

Contrast with Team/SSC approaches (3 of 7)

- SSC's approach:
  - All steps are the same as in the Team's approach, except for step 2, which is replaced by the following:
  - Do not average the model-specific $F_{40\%}$ and $B_{40\%}$ reference points as in step 2 of the Team's approach, but instead use the averages computed in step 1 to compute an "average" $F_{40\%}$ value; then compute the average of the model-specific mean recruitments and use that average along with the averages computed in step 1 to compute an "average" $B_{40\%}$ value.

Contrast with Team/SSC approaches (4 of 7)

- Legend: blue = Model 2, orange = SSC, gray = GPT, yellow = Model 1.
- Team maxABC = 18.974, SSC maxABC = 41.577.



Contrast with Team/SSC approaches (5 of 7)

- The problem of nonlinearity:
  - Both the harvest control rule, and the models themselves, result in *abc* values that are nonlinear transforms of the parameters that are actually involved in minimizing the objective function.
    - Note that biomass is *not* one of the "basic estimates from the assessment;" it is a *function* of the estimated parameters.
  - By analogy, which is the better way to estimate average weight:
    - average the weights of the fish in the sample, or
    - fit a weight-at-length model, then average the lengths of the fish in the sample, then insert that average length into the model?
- (continued on next slide)

Contrast with Team/SSC approaches (6 of 7)

- The problem of nonlinearity (continued):
- As was stressed repeatedly at last year's Team workshop on model averaging and *abc* reductions, it is impossible to produce an "internally consistent" ensemble when nonlinearities are present
- Note the following Team recommendation (9/18, SSC endorsed 10/18):
  - "Assuming that some sort of model averaging is involved, an ensemble model should be treated the same as any other model (i.e., an ensemble is a 'model' and should be treated as such in reference to the existing language in the FMP and SAFE report guidelines)."
  - That is, rather than trying to reverse-engineer a single model that matches the behavior of the ensemble, the ensemble itself should be treated the same as any other model.
- (continued on next slide)

Contrast with Team/SSC approaches (7 of 7)

- The problem of nonlinearity (continued):
  - The solution is simple:
    - If an optimal estimate of $F_{40\%}$ is desired, compute the ensemble estimate of $F_{40\%}$.
    - If an optimal estimate of current biomass is desired, compute the ensemble estimate of current biomass.
    - If an optimal estimate of the *ofl* distribution is desired, compute the ensemble estimate of the *ofl* distribution.
    - Etc.
  - The set of resulting estimates will not map into any single model, but they *will* be consistent with the Team/SSC advice to treat the ensemble as a model, and they *will* be optimal!

Wrap-up (both presentations)

Feedback needed from Team

- Which models to include in final assessment?
- Pursue model averaging in final assessment?
  - If so:
    - How to incorporate “justifiable” and “plausibility” weightings?
    - How to calculate ensemble harvest specs?
- Guidance for sampling of State-managed fishery?