Model averaging by cross-conditional decision analysis

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

Abstract

This paper introduces cross-conditional decision analysis (CCDA), a Bayesian method of model averaging that incorporates, simultaneously, both the probability of alternative models being "true" and the relative performance of each model when one of the other models is the "true" one. Importantly, CCDA is applicable to cases in which true values of the quantity of interest are not available, which would ordinarily preclude "training" of the ensemble. CCDA circumvents this difficulty by treating each model in the ensemble, one at a time, *as if* it were true, then computing the risk (expected loss) 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), then multiplying the results by the probability that the pivot model is the true model, then summing across pivot models to obtain the risk for the entire ensemble, then tuning the performance weights so as to minimize the ensemble risk, then, finally, using those weights to create an ensemble probability mass function from which the optimal value of the quantity of interest can be estimated. The suggested loss function allows results to be tuned to any desired level of risk aversion.

Introduction

The need to account for uncertainty when providing fishery management advice has been acknowledged countless times in the last four decades or so. Early methods for doing so tended to focus on the uncertainty associated with estimates obtained by a single model (e.g., Walters 1975), but uncertainty associated with model structure has also been addressed, with over 20 papers now having presented applications of model averaging in a fishery assessment or management context. Most often, treatments of uncertainty have followed a Bayesian approach, at least to some extent, with the authors frequently referring to the process as a "Bayesian decision analysis," or words to that effect. Within the model averaging context in particular, the Bayesian perspective has been a fairly consistent feature in the fishery literature ever since Sainsbury (1988).

A note on use of the term "Bayesian decision analysis" is warranted at this point. All authors invoking this term require, at a minimum, use of a Bayesian posterior distribution of the quantity of interest, or at least an approximation thereof. However, beyond this minimum requirement, some disagreement exists. Some authors require nothing beyond the posterior distribution in order to complete the analysis. These authors typically define "risk" as cumulative probability, as calculated from the posterior distribution (e.g., the "$p*$" approach of Shertzer et al. 2008). However, other authors require, in addition to the posterior distribution, specification of a loss function in order to complete the analysis. These authors typically define risk as the expected loss; that is, the integral of the product of the posterior distribution (or approximation thereof) and the loss function (e.g., Thompson 1992). The distinction will be addressed further in the Discussion, but it should suffice for now to say that the latter interpretation will be adopted here.

This information is distributed solely for the purpose of pre-dissemination peer review under applicable information quality guidelines. It has not been formally disseminated by the National Marine Fisheries Service and should not be construed to represent any agency determination or policy.

The basic steps in a Bayesian analysis are thus as follow:

1. Choose a quantity of interest
2. Calculate the posterior distribution of the quantity of interest
3. Choose a loss function
4. Integrate the product of the posterior distribution and the loss function
5. Minimize the integral (i.e., the risk)

In the context of marine fishery management in the U.S., step 1 in the above list might result in selection of the "overfishing level" (*ofl*) as the quantity of interest, which is typically defined by a harvest control rule that is related to the fishing mortality rate corresponding to maximum sustainable yield (*msy*). In the simplest case, the harvest control rule sets *ofl* for the coming year equal to the yield that would be achieved by fishing at the rate corresponding to *msy*, while more complicated harvest control rules vary the fishing mortality rate as a function of spawning biomass or other measure of reproductive output.

Step 2 in the above list becomes complicated when dealing with an ensemble (i.e., a set of models), because the posterior distribution (probability density function (pdf) in the case of a continuous random variable or probability mass function (pmf) in the case of a discrete random variable) will be an average of the *model-specific* posterior distributions, but there is yet no consensus on how this average should be computed; that is, whether the average should be unweighted or weighted and, if the latter, how the weights should be specified.

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 (e.g., Hill et al. 2007, Ianelli et al. 2016). Although some studies of ensembles have successfully produced fully Bayesian probabilities for the models in the ensemble (e.g., Sainsbury 1988, Patterson 1999, Brandon and Wade 2006), most authors have defaulted to approximations such as purely subjective "plausibility weighting" (e.g., Butterworth et al. 1996) or weights based on importance sampling (e.g., McAllister and Kirchner 2002), harmonic mean approximation (e.g., Parma 2002, Millar et al. 2015), Akaike Information Criterion (AIC; e.g., Millar et al. 2015, Rossi et al. 2019), Bayesian (Schwarz) Information Criterion (BIC; e.g., Brodziak and Legault 2005), Deviance Information Criterion (DIC; e.g., Wilberg and Bence 2008), bootstrapping (e.g., Millar et al. 2015) cross-validation (e.g., Scott et al. 2016, Rossi et al. 2019), or retrospective analysis (e.g., Rossi et al. 2019). Equal weighting of models has also been used (e.g., Stewart and Martell 2015, Ianelli et al. 2016, Rossi et al. 2019).

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.

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. Similar problems exist for many other quantities estimated in stock assessments.

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.

Instead, the method developed here treats each model in the ensemble, one at a time, *as if* it were true, then computes 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), then multiplies the results by the probability that the pivot model is the true model, then sums across pivot models to obtain the risk for the entire ensemble, then tunes the weights so as to minimize the ensemble risk, then, finally, uses those weights to create an ensemble pmf from which the optimal value of the quantity of interest can be estimated.

The above process provides a *cross-conditional decision analysis* (CCDA), the specific steps of which are detailed more explicitly in the next section.

Methods

The following notational conventions are used:

- Capitalization:
  - Names of matrices consist of, or begin with, an upper-case letter.
  - Names of vectors and scalars consist of, or begin with, a lower-case letter.
- Font:
  - Names of matrices and vectors appear in bold font.
  - Names of scalars appear in italicized font.
- The notation $\mathbf{Z}^{\langle j \rangle}$ represents column *j* of matrix **Z**.

The following loss function, which is described in detail in Appendix A, will be 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 "negative" risk aversion (i.e., risk proclivity). Here, risk aversion means that any underestimate is preferred to an overestimate of the same magnitude.

The following procedure is fairly general, and should be applicable to a wide range of choices as to the quantity of interest, with two constraints: 1) the quantity of interest cannot take negative values; and 2) 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 given above.

As shown in Appendix A, 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 ($g_y(y)$).

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

For an ensemble containing *nmod* candidate models, model averaging by cross-conditional decision analysis consists of the following steps:

1. Choose a value for *ra*.
2. For each model i = 1,2,...,*nmod* (referred to below as the "pivot" model, indexed):
    a. Fit pivot model i to the data; this is the "base" run of the pivot model.
    b. Using the parameters estimated from the base run, generate *nsim* sets of conditional parametric bootstrap data.
    c. Fit pivot model i to each bootstrap data set k=1,2,...,*nsim*, resulting in a set of *nsim* estimates of y ($\mathbf{yest}_i$), which will be taken to characterize the distribution of y, conditional on the structure of the pivot model being "true."
    d. Compute the risk-minimizing value of $\hat{y}$ ($yopt_i$) conditional on the structure of the pivot model being "true," which will be the $\mathbf{yest}_i$ mean of order $1-ra$.
    e. Fit each model $j \neq i$ in the ensemble to each of the *nsim* sets of bootstrap data generated by pivot model i, resulting in a vector $\mathbf{yest}_j$ for each such model, which, together with $\mathbf{yest}_i$, form the columns of the matrix $\mathbf{Yest}_i$ (note: after steps 2a-2e have been completed for all pivot models, a total of *nmod* **Yest** matrices will have been created, one for each pivot model, and each **Yest** matrix will consist of *nsim* rows and *nmod* columns).
    f. For each set of bootstrap data k=1,2,...,*nsim*, if *any* of the *nmod* fitted models fails to produce a positive definite Hessian matrix or if the maximum gradient exceeds a specified tolerance (e.g., 0.01), delete the corresponding row from $\mathbf{Yest}_i$, resulting in a matrix $\mathbf{Yest\_use}_i$.
    g. Determine the probability (p_i) that the structure pivot model i corresponds to the structure of the true model, using either quantitative or qualitative methods.
3. Create a vector of weights **w**, where each element $0 \leq w_i \leq 1$, i = 1,2,..,*nmod*, and the vector is constrained to sum to unity; set equal initially to the vector of probabilities **p**.
4. Define a conditional risk for each pivot model i=1,2,...,*nmod* as follows:

$$condrisk(\mathbf{w})_i = \left(\frac{1}{nuse_i}\right) \sum_{k=1}^{nuse_i} loss\left(\sum_{j=1}^{nmod} \left(w_j (\mathbf{Yest\_use}_i)_{k,j}\right) \mid yopt_i, ra \right),$$

5. Define the overall risk (i.e., expected loss) associated with **w** as

$$risk(\mathbf{w}) = \sum_{i=1}^{nmod} \left(p_i condrisk(\mathbf{w})_i\right).$$

6. Minimize *risk*(**w**) w.r.t. **w**.
7. Form a set of *nbin*=4 equal-sized histogram bins spanning the range from min(**Yest_use**) to max(**Yest_use**); note that the number of bins will be adjusted later.
8. For each combination of pivot model i=1,2,...,*nmod* and candidate model j=1,2,...,*nmod*, form a histogram from $\mathbf{Yest\_use}_i^{\langle j \rangle}$.
9. Form an *nmod*×*nmod* matrix of probability mass functions (**Pmf**) by converting each histogram into a probability mass function by normalizing so that the sum of the bar heights equals unity.
10. Form a weighted average probability mass function (*pmf*) across models by computing the probability in each *bin*=1,2,...,*nbin* follows:

$$pmf_{bin} = \sum_{i=1}^{nmod} p_i \sum_{j=1}^{nmod} w_j \left(Pmf_{i,j}\right)_{bin} .$$

11. Form a weighted average cumulative mass function (*cmf*) by computing the cumulative probability in each *bin*=1,2,...,*nbin* as follows:

$$cmf_{bin} = \sum_{ibin=1}^{bin} pmf_{ibin} .$$

12. Compute the median of the distribution by interpolating linearly between cmf_{binlo} and cmf_{binhi}, where *binlo* is the largest value of *bin* such that cmf_{bin}<0.5 and *binhi* is the smallest value of *bin* such that cmf_{bin}>0.5.
13. Double the number of equal-sized histogram bins, return to step 8, and repeat until the median computed in step 12 does not change by more than a specified tolerance (e.g., 0.001).
14. Compute the overall risk-minimizing value of $\hat{y}$ (as opposed to the risk-minimizing value of $\hat{y}$ computed for each pivot model *i*, $yopt_i$, in step 2d), which will be the mean of order 1−*ra*, based on the probability mass function computed in step 10.
15. Optional: Perform a series of 10-fold cross validations to estimate the distribution of **w** estimates and to explore the distribution of risk when CCDA is applied to data not included in the estimation of **w** (i.e., the out-of-sample performance).

As a test case, CCDA was applied to an ensemble of simple surplus production models, with *ofl* chosen as the quantity of interest. Full details are described in Appendix B, but 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.

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.
  - Model 6 included covariates #1 and #3.
  - Model 7 included covariates #2 and #3.
- Model 8 included all three environmental covariates.

Model 5 was chosen as the “true” model, and so was used as the operating model that generated the data set to which the base run of each model was fit. Once the base run for each model was made, *nsim*=500 sets of data were generated by a parametric bootstrap, conditional on the respective pivot model.

The probability of model *i* being true was a linear function of the number of environmental covariates used in that model ($nvar_i$, as distinguished from *nv*=max(**nvar**)):

$$p_i = (5 + nvar_i)\left(5nmod + \sum_{j=1}^{nmod} nvar_j\right)^{-1} .$$

The decision analysis was conducted twice, once with *ra*=0, representing risk neutrality, and once with *ra*=2, representing a risk-averse alternative.

Results

Figure 1 shows some of the results of the base run for each model, giving estimates of the msy exploitation rate (*umsy*), the projection year stock size (*xpro*), and the projection year *ofl*; all relative to the respective true value.

Out of 4000 bootstrap data sets (8 models × 500 bootstrap data sets per model), all models were determined to have converged in 2594 (64.8% of all runs). The number of usable runs for each pivot model were as follow:

Model:	1	2	3	4	5	6	7	8
Runs:	265	307	340	334	361	295	352	340

The complete set of 8×8 histograms of *oflest_use* are shown in Appendix C, with one figure per pivot model and results (histograms) from eight candidate models per pivot model.

The value of *yopt* (i.e., the risk-minimizing value of *y* for a given model when fit to the bootstrap data sets generated from its own base run) for each model 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

The values of *yopt* are higher than the base value in all instances except one (Model 1, with *ra*=2), and the value of *yopt* from the risk-averse case is closer to the base value than the value of *yopt* from the risk-neutral case in all instances.

A matrix of cross-conditional risks (**CCR**), independent of both **p** and **w**, was computed first by computing nominal values for each pivot model *i*=1,2,...,*nmod* and candidate model *j*=1,2,...,*nmod* as

$$ccr_{i,j} = \left(\frac{1}{nuse_i}\right)\sum_{k=1}^{nuse_i} loss\big((\mathbf{Yest\_use}_i)_{k,j} \,|\, yopt_i, ra\big) ,$$

and then normalizing so that the values sum to unity, giving a matrix of cross conditional *relative* risks (**CCRR**):

$$CCRR = CCR\left(\sum_{i=1}^{nmod}\sum_{j=1}^{nmod} ccr_{i,j}\right)^{-1}$$

The results of the above calculations are shown in Table 1. For the risk-neutral (*ra*=0) case, the riskiest combinations tended to occur when Model 1 was the pivot model or when Model 3 was the candidate model, as indicated by cells shaded toward the green end of the red-green scale. For the risk-averse (*ra*=2) case, the riskiest combinations tended to occur when either Model 4 or Model 6 was the pivot model or when Model 3 or Model 7 was the candidate model. (Note that results cannot be compared between the two halves of Table 1, as only results for a common value of *ra* are comparable.)

Correlations between model results may also be of interest. The full set of cross-conditional correlations between candidate models and pivot models is shown in Table 2. The vast majority of correlations were positive, with only 46 out of a possible 224 (about 21%) being negative. The largest (in absolute value) negative correlation was –0.289.

The mean off-diagonal correlations between models were as follow:

Model:	1	2	3	4	5	6	7	8
Corr.:	0.516	0.461	0.405	0.437	0.348	0.441	0.352	0.361

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.

Table 3 shows the conditional risks (step 4 in the algorithm described in the Methods section) in both absolute and relative terms for the risk-averse (*ra*=0) and risk-neutral (*ra*=2) cases, and, for each case, under both equal weighting (i.e., where the elements of **w** were all set equal to 1/*nmod*, but **p** was unchanged) and optimal weighting. Under all case×weighting combinations, Model 7 contributes the least to the overall risk. Model 6 contributes the most in the risk-neutral case under both equal weighting and optimal weighting, while Model 3 contributes the most in the risk averse case under both equal weighting and optimal weighting.

An alternative way to express the relative risk associated with a given model i is to compute a "risk ratio," in which the numerator consists of the expected loss under $w_i = 1$ and $w_j = 0$ for all $j \neq i$ and the denominator consists of the expected loss under optimal weighting (**p** is unchanged).

The risk ratios, expressed on a log (base 10) scale, were as follow (the last column shows another alternative risk ratio in which the numerator consists of the expected loss under **w**=**p**):

ra	1	2	3	4	5	6	7	8	**w**=**p**
0	0.224	0.450	0.316	0.368	0.164	0.351	0.434	0.220	0.026
2	0.205	0.281	7.523	0.242	7.043	0.260	7.557	7.181	0.037

Note that assigning all weight to Model 5, which is the true model, results in a large increase in expected loss in the risk-averse (*ra*=2) case.

Table 4 shows various statistics of the distributions for the risk-neutral (*ra*=0) and risk-averse (*ra*=2) cases. The values in this table are based on an *nbin* value of 2048, as the median values showed no changes out to three significant digits between *nbin* values of 1024 and 2048. The optimal value of *y* for the risk-averse case (0.250) is about 10% less than the optimal value for the risk-neutral case (0.278). The final row in the table shows the cumulative probability (*p**) associated with the optimal value in the respective distribution. 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.

The large ratio between the number of bins and the number of usable *ofl* estimates means that *pmf* is quite noisy, so it was smoothed by averaging across the nearest *nsmooth* points (in both directions, to the extent possible), where *nsmooth* was set at the largest value that resulted in each of the first three non-central moments of the smoothed distribution diverging from the unsmoothed distribution by a relative value of no more than 0.005, giving *nsmooth* values of 26 and 24 for the risk-neutral and risk-averse cases, respectively. The smoothed *pmf*s are shown in Figure 2a, and both the smoothed and unsmoothed *cmf*s are shown in Figure 2b. In Figure 2b, the almost exact matches between the solid green and dashed red curves and between the solid black and dashed yellow curves shows that the amount of smoothing was not excessive. The vertical dashed lines in Figure 2 show the locations of the respective optima for the risk-neutral (green) and risk-averse (black) cases. Although most of the statistics in Table 4 appear very similar between the two cases, Figure 2 shows that the shapes of the distributions are quite distinct. In particular, the risk-averse *pmf* shows clear bimodality.

Figure 3 shows smoothed probability mass functions for the individual models in the risk-neutral (Figure 3a) and risk-averse (Figure 3b) cases, unweighted by their respective probabilities, while Figure 4 shows the analogous results when weighted by the respective probabilities. It is clear that Models 5, 7, and 8 contribute most to the bimodality exhibited by the overall *pmf* in the risk-averse case.

Table 5 shows the results of the cross-validation exercise, in which the 10-fold cross-validation was repeated 10 times, with the folds chosen randomly each time. Results are shown with respect to model weights in Table 5a, and with respect to expected loss in Table 5b. The 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. 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.

Discussion

Using CCDA to produce harvest specifications

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, the first of which involves *nmod* individual optimizations:

- 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. In general terms, the former would be a natural choice for a “limit” harvest amount, while the latter would be a natural choice for a “target” reference amount.

More particularly, in the context of U.S. marine fishery management, the former would be a natural choice for "the" *ofl* (i.e., the *ofl* value specified in regulations), while the latter would a natural choice for the "acceptable biological catch" (*abc*), defined in Federal guidelines (https://federalregister.gov/d/2016-24500) 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 "Council's 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 associated with the respective optimal values.

Some previous applications of Bayesian decision analysis (e.g., Thompson 1992, 1996, and 1999; also section 3.1 of Restrepo et al. 1998), have cast the distinction between limit and target harvest rates differently, with a focus on determining which reference points should be used to *define* the limit and target control rules, whereas here the focus is on optimizing the catches resulting from *estimates* of those control rule reference points (and any other parameters that are needed to estimate those catches). See also Appendix A.

What constitutes a Bayesian decision analysis?

As noted in the Introduction, whether in a single-model or ensemble context, the fishery assessment and management literature evidences some disagreement about the necessary elements of a Bayesian decision analysis. Many studies that invoke this label use it to mean simply that a Bayesian posterior will be the end product, with the "decision analysis" consisting of finding the value of the quantity of interest that corresponds to a specified cumulative probability (e.g., the *p** approach of Shertzer et al. 2008, which has become something of a standard in U.S. marine fishery management). The approach taken here, on the other hand, follows authors such as DeGroot (1970), who require that the distribution be used together with a loss function to solve for the risk-minimizing decision (or, equivalently, use of the distribution together with a utility function to solve for the decision that maximizes expected utility).

Given the loss function used here, the two approaches are similar to the extent that the "decision" can be computed directly from the distribution (the value that corresponds to a critical percentile in the former case; an order mean in the latter). They are also similar in that both require specification of a decision criterion (the critical percentile in the former; the level of risk aversion in the latter). They are dissimilar in that the former is based on satisfying a constraint, whereas the latter is based on optimization. They are also dissimilar in that the former considers only the question of whether $\hat{y} > y$, without regard to the amount by which $\hat{y}$ might differ from y or the consequences of such errors, in contrast to the latter.

One way to interpret the difference between the two approaches is that, although both use Bayesian posterior distributions, the former uses them in an essentially "frequentist" manner to generate a decision, whereas the latter maintains the Bayesian perspective all the way through the decision-making process.

Bootstrap distributions as an approximation of Bayesian posterior distributions

As CCDA has been implemented so far, the distribution of the quantity of interest for each pivot model is based on a set of conditional parametric bootstraps. Use of bootstrap distributions has a long history in fishery science, going back at least as far as Deriso et al. (1985) and Kimura and Balsiger (1985). However, bootstrap distributions are only approximations to Bayesian posterior distributions. Therefore, if a method can be found for obtaining the Bayesian posterior distributions needed by CCDA, it would probably be preferable to use such an approach (subject to computational feasibility), but this appears difficult, given the need to generate a parallel set of distributions for each candidate model. Whether bootstrapped distributions are sufficiently good approximations of Bayesian posterior distributions remains an open question, with many studies having been conducted, often with divergent conclusions.

Among the studies that have evaluated the performance of alternative uncertainty estimators (Bayesian posterior distribution, various types of bootstrap distributions, Hessian matrix inversion, delta method, frequentist methods, likelihood methods, and MCMC) are those by Mohn (1993, 2009), Punt and Butterworth (1993), Gavaris (1999), Patterson (1999), Gavaris et al. (2000), Restrepo et al. (2000), Patterson et al. (2001), Zhou (2002), Magnusson et al. (2013), and Elvarsson et al. (2014).

Differences in results have been attributed (see, for example, Magnusson et al. 2013) to the specific type of bootstrap being evaluated (e.g., parametric versus non-parametric, bias-corrected versus not), the performance measure being used (full pdfs versus confidence intervals or variances), the overall approach (empirical versus simulation-based), and, in the case of simulation-based approaches, the complexity of the operating model. Another possible issue is the effect of assuming the wrong functional form for the likelihood when generating the posterior distribution.

In their review of methods for model averaging, Millar et al. (2015) considered weights based on bootstrapping, but focused primarily non-parametric bootstraps, which they noted are often not well suited to fishery applications. They suggested that parametric bootstraps might be a better alternative, but cautioned that the operating model underlying the parametric bootstraps might cause over-weighting of those models whose structures were the most similar to the operating model. However, the cross-conditioning aspect of CCDA addresses such potential over-weighting explicitly, by requiring each model to take a turn as the operating model.

Given the facts that: 1) all fishery modeling involves approximations, bootstrap approximations to Bayesian posterior distributions remain a contender after extensive analysis, and 3) the potential shortcoming noted by Millar et al. (2015) has been addressed; it is reasonable to conclude that their use in CCDA is likely not a critical flaw.

Assuming that the ensemble contains the true model

The assumption that the ensemble contains the true model, while almost surely invalid in the strict sense, is widely made in the model selection literature, particularly that portion of the literature that advocates use of BIC (e.g., Bernardo and Smith (1994), Kadane and Lazar (2004), Chaurasia and Harel (2013), Aho et al. (2014)). Another consideration is that, even when the impossibility of ever identifying the true model is acknowledged, the fact remains that, in practice, the setting of harvest specifications typically proceeds *as though* one of the models in the ensemble is the true model (e.g., Kass and Raftery 1995).

It should be emphasized that the reason for making this assumption here is that it is needed to generate true data on which the ensemble can be trained. The assumption that the ensemble contains the true model should therefore be viewed as a matter of convenience (or even necessity) rather than ontology.

Practical considerations

1. A significant practical consideration for implementing CCDA here is the amount of time required to conduct the analysis. For example, in the application presented here, an ensemble of *nmod*=8 models was involved, each of which was used to generate *nsim*=500 bootstrap data sets, each of which had to be fit by each candidate model, resulting in a total of *nmod*×*nsim*×*nmod*=32,000 model runs to conduct a single CCDA. Even though the models in the ensemble involved only 5-8 estimated parameters (Appendix B) and all steps were fully automated, performing all of the model runs still took at least a couple of days. Of course, use of smaller values for either *nmod* or *nsim* might be acceptable, but if the ensemble involves very complicated models, application of CCDA could potentially take a long time. In the event that the turn-around time for a stock assessment (i.e., the interval between the time when the new data become available and the time when the assessment must be completed) is too short to permit a full CCDA, a reasonable compromise might be to estimate the performance weights on the basis of the data used in the previous assessment and assume that they are unchanged for the current assessment (of course, this would require that the ensemble not change between assessments).

2. Two other practical considerations arise from subtleties involved with step 2b in the algorithm (see Methods): 2a) The resulting bootstrap data will truly be "conditional" only if the parameters estimated by the pivot model include one or more parameters describing the distribution of one or more data sets (e.g., if the measurement error variance for a particular data set is estimated by the pivot model). For any data set that does not involve any parameters that are estimated by any of the models, a single set of *nsim* bootstrap data sets can be generated and used for all models, thus saving at least a little time in applying CCDA. 2b) It may be the case that some model "A" might estimate the measurement error variance for some data set "B," while model "C" might not use data set "B" at all, meaning that, when model C takes its turn as the pivot model, not only will it not be possible to generate bootstrap values for data set B that are truly conditional on model C being the true model, it is not even obvious how the suggestion presented for subtlety 2a would be applied. A straightforward solution would be to generate the bootstrap data from the parameter values estimated on the basis of the sampling design, which are typically available for the data sets most commonly used in fishery stock assessments. In the event that estimating the parameter values on the basis of the sampling design is impossible, it might be necessary to remove either model A or model B from the ensemble, but this circumstance does not appear likely.

3. A third practical consideration relates to the difference between the two main interpretations of Bayesian decision analysis discussed above, where one takes an essentially frequentist approach in deriving a decision from the posterior distribution and the other maintains the Bayesian perspective all the way through the decision-making process, which is that the optimal estimate of *ofl* may in some cases exceed the median of the optimized *ofl* distribution. This occurred in the example presented here (Table 4), where the optimal *ofl* given *ra*=0 occurred at the 63rd percentile of the corresponding distribution and the optimal *ofl* given *ra*=2 was barely above the median of the corresponding distribution (but below the median of the distribution given *ra*=0). In the frequentist interpretation of the posterior distribution, these results are concerning, particularly for the optimal *ofl* given *ra*=0, because, in the frequentist interpretation, the only thing that matters is the probability of being above or below the true-but-unknown value of *ofl*; how *much* above or below is irrelevant. Therefore, in the frequentist interpretation, any estimate of *ofl* that lies above the median appears "risky," even if it is the optimal risk-neutral (or risk-averse) estimate. In the fully Bayesian interpretation, on the other hand, the percentile of the distribution is irrelevant, because the *entire distribution* (not just the median, or any other single percentile) has been factored systematically into the calculation of the optimum, including, importantly, the relative undesirability of each possible over- or under-estimate.

One fact that might give added weight to this (third) practical consideration, at least in the context of U.S. marine fishery management, is the U.S. Court of Appeals (2000) decision in the case of the fishery for summer flounder (*Paralichthys dentatus*) off the northeastern coast of the U.S. (Tercerio 2002). There, the court ruled that a harvest limit in excess of the median was impermissible, based on "the at-least-50% likelihood required by statute and regulation." However, the statute and regulation to which the court referred are unclear. For example, the Magnuson-Stevens Fishery Conservation and Magnuson Act (MSFCMA) makes no mention of such a requirement, and neither did the version of the Federal guidelines for National Standard 1 of the MSFCMA that was current at the time of the ruling (at least not in the context of setting a harvest limit). Moreover, the current version of those guidelines (https://federalregister.gov/d/2016-24500) explicitly renders the 50% standard optional (§600.305(f)(2)(i)) and, when introducing the changes that were contemplated when developing the current version (https://www.federalregister.gov/documents/2015/01/20/2015-00586/magnuson-stevens-act-provisions-national-standard-guidelines), the National Marine Fisheries Service explicitly cited "expressed interest in using a decision theoretic approach" as one of the reasons for making the 50%

standard optional. Thus, the extent to which the summer flounder ruling continues to be a constraint on admissible methods for producing harvest specifications is at least questionable. In the event that the ruling is interpreted as constituting a continuing constraint, the optimal values produced by CCDA would have to be supplemented by language to the effect, "or the median of the corresponding distribution, whichever is less."

4. A fourth practical consideration is the need to estimate each model's probability of being "true." While quantitative estimation of these probabilities would clearly be desirable, it appears likely that qualitative estimation based on purely subjective evaluations of relative "plausibility" will be more common. Butterworth et al. (1996) provide one possible set of guidelines for such evaluations. In the event that it is simply impossible to afford any single model more probability of being "true" than any other model, then assigning equal probability to all models would be the obvious course of action.

5. A final practical consideration is the need to choose one or more *ra* values. If the objective is to produce a risk-neutral optimum, setting *ra*=0 is a straightforward choice. However, if the choice is to produce a risk-averse or risk-prone optimum, it will be necessary to choose some $ra \neq 0$. If the management system already contains both limit and target harvest control rules, it may be possible to reverse-engineer the *ra* value that is implicit in the existing target harvest control rule. Otherwise, it may be necessary to simulate a large number of examples across a range of *ra* values and a range of data imprecision and ask fishery managers to select the *ra* value that captures the attitude toward risk that is most appropriate for managing the fisheries under their jurisdiction. Although selecting an appropriate *ra* value is a nontrivial problem, surely it is no more difficult than selecting an appropriate *p** value. Moreover, because the fully Bayesian approach considers both the magnitudes and relative undesirability of possible estimation errors, as opposed to considering only the probability that a positive error will occur (e.g., an *ofl* estimate higher than the true-but-unknown value), *ra* may actually be more meaningful to fishery managers than *p**.

Conclusion

CCDA provides a solution to the tension between model weighting based on posterior probability and model weighting based on predictive performance. Importantly, CCDA is applicable to cases in which true values of the quantity of interest are not available, which would ordinarily preclude "training" of an ensemble. By taking a fully Bayesian approach, CCDA allows the same risk attitude to be incorporated throughout the entire decision analysis, and so produces a fully integrated and highly coherent set of results. By performing separate CCDAs for alternative levels of risk aversion, the implications of alternative risk attitudes can be explored and used to make management recommendations. However, the substantial time requirements of CCDA may pose challenges for stock assessments with a short turn-around time.

References

Aho, K., D. Derryberry, and T. Peterson. 2014. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95(3):631-636.

Anderson, S. C., A. B. Cooper, O. P. Jensen, C. Minto, J. T. Thorson, J. C. Walsh, J. Afflerbach, M. Dicky-Collas, K. M. Kleisner, C. Longo, G. Chato Osio, D. Ovando, I. Mosqueira, A. A. Rosenber, and E. R. Selig. 2017. Improving estimates of population status and trend with superensemble models. *Fish Fish.* 18:732-741.

Bernardo, J. M., and A. F. M. Smith. 1994. *Bayesian theory*. John Wiley and Sons, Ltd (Chichester, West Sussex, England). 586 p.

Brandon, J. R., and P. R. Wade. 2006. Assessment of the Bering-Chuckchi-Beaufort Seas stock of bowhead whales using Bayesian model averaging. *J. Cetacean Res. Manage*. 8:225-239.

Brodziak, J., and C. M. Legault. 2005. Model averaging to estimate rebuilding targets for overfished stocks. *Can. J. Fish. Aquat. Sci*. 62:544-562.

Butterworth, D. S., A. E. Punt, and A. D. M. Smith. 1996. On plausible hypotheses and their weighting, with implications for selection between variants of the Revised Management Procedure. *Rep. int. Whal. Comm.* 46:637–640.

Chaurasia, A. and O. Harel. 2013. Model selection rates of information based criteria. *Electron. J. Stat.* 7:2762-2793.

DeGroot, M. H. 1970. *Optimal statistical decisions*. McGraw-Hill, New York, 489 p.

Deriso, R. B., T. J. Quinn II, and P. R. Neal. 1985. Catch-age analysis with auxilliary information. *Can. J. Fish. Aquat. Sci*. 42:815-824.

Elvarsson, B., L. Taylor, V. M. Trenkel, V. Kupca, and G. Stefansson. 2014. A bootstrap method for estimating bias and variance in statistical fisheries modelling frameworks using highly disparate datasets. *Afr. J. Mar. Sci.* 36:99-110.

Gavaris, S. 1999. Dealing with bias in estimating uncertainty and risk. *In* V. R. Restrepo (ed.), *Proceedings of the 5th National NMFS Stock Assessment Workshop: Providing scientific advice to implement the precautionary approach under the Magnuson-Stevens Fishery Conservation and Management Act*, p. 46-50. NOAA Tech. Memo. NMFS-F/SPO-40. National Marine Fisheries Service, NOAA. 1315 East-West Highway, Silver Spring, MD 20910.

Gavaris, S., K. R. Patterson, C. D. Darby, P. Lewy, B. Mesnil, A. E. Punt, R. M. Cook, L. T. Kell, C. M. O'Brien, V. R. Restrepo, D. W. Skagen, and G. Stefánsson. 2000. Comparison of uncertainty estimates in the short term using real data. ICES CM 2000/V:03.

Hill, S. L., G. M. Watters, A. E. Punt, M. K. McAllister, C. Le Quéré, and J. Turner. 2007. Model uncertainty in the ecosystem approach to fisheries. *Fish Fish*. 8:315-336.

Ianelli, J., K. K. Holsman, A. E. Punt, and K. Aydin. 2016. Multi-model inference for incorporating trophic and climate uncertainty into stock assessments. *Deep-Sea Res. Pt. II* 134:379-389.

Kadane, J. B., and N. A. Lazar. 2004. Methods and criteria for model selection. *J. Am. Stat. Assoc.* 99(465):279-290.

Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc*. 90:773-793.

Kimura, D. K., and J. W. Balsiger. 1985. Bootstrap methods for evaluating sablefish pot index surveys. *N. Am. J. Fish. Manage*. 5:45-56.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran. 1999. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* 285:1548-1550.

Magnusson, A., A. E. Punt, and R. Hilborn. 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. *Fish Fish.* 14:325-342.

McAllister, M., and C. Kirchner. 2002. Accounting for structural uncertainty to facilitate precautionary fishery management: illustration with Namibian orange roughey. *Bull. Mar. Sci*. 70:499-540.

Millar, C. P., E. Jardim, F. Scott, G. Chato Osio, I. Mosqueira, and N. Alzorriz. 2015. Model averaging to streamline the stock assessment process. *ICES J. Mar. Sci*. 72:93-98.

Mohn, R. K. 1993. Bootstrap estimates of ADAPT parameters, their projection in risk analysis and their retrospective patterns. In Smith, S. J., J. J. Hunt, and D. Rivard, eds. Risk evaluation and biological reference points for fisheries management. *Can. Spec. Publ. Fish. Aquat. Sci*. 120:173-184.

Mohn, R. 2009. The uncertain future of assessment uncertainty. In R. J. Beamish and B. J. Rothschild, eds. *The Future of Fisheries Science in North America*. Fish and Fisheries Series, Springer Science + Business Media, p. 495-504.

Parma, A. M. 2002. Bayesian approaches to the analysis of uncertainty in the stock assessment of Pacific halibut. In J. M. Berkson, L. L. Kline and D. J. Orth, eds. Incorporating uncertainty into fishery models. *Amer. Fish. Soc. Symp*. 27:113-136.

Patterson, K. R. 1991. Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. *Can. J. Fish. Aquat. Sci.* 56:208-221.

Patterson, K. R. 1999. Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Can. J. Fish. Aquat. Sci. 56:208-221.

Patterson, K. R., R. Cook, C. Darby, S. Gavaris, L. Kell, P. Lewy, B. Mesnil, A. Punt, V. Restrepo, D. W. Skagen, and G. Stefánsson. 2001. Estimating uncertainty in fish stock assessment and forecasting. *Fish Fish*. 2:125-157.

Punt, A. E., and D. S. Butterworth. 1993. Variance estimates for fisheries assessment: their importance and how best to evaluate them. In Smith, S. J., J. J. Hunt, and D. Rivard, eds. Risk evaluation and biological reference points for fisheries management. *Can. Spec. Publ. Fish. Aquat. Sci*. 120:145-162.

Restrepo, V. R., K. R. Patterson, C. D. Darby, S. Gavaris, L. T. Kell, P. Lewy, B. Mesnil, A. E. Punt, R. M. Cook, C. M. O'Brien, D. W. Skagen, and G. Stefánsson. 2000. Do different methods provide accurate probability statements in the short term? ICES CM 2000:/V:08.

Restrepo, V. R., G. G. Thompson, P. M. Mace, W. L. Gabriel, L. L. Low, A. D. MacCall, R. D. Methot, J. E. Powers, B. L. Taylor, P. R. Wade, and J. F. Witzig. 1998. *Technical guidance on the use of precautionary approaches to implementing National Standard 1 of the Magnuson-Stevens Fishery Conservation and Management Act*. NOAA Tech. Memo. NMFS-F/SPO-31. National Marine Fisheries Service, NOAA. 1315 East-West Highway, Silver Spring, MD 20910. 54 p.

Rosenberg, A. A., K. M. Kleisner, J. Afflerbach, S. C. Anderson, M. Dickey-Collas, A. B. Cooper, M. J. Fogarty, E. A. Fulton, N. L. Gutierrez, K. J. W. Hyde, E. Jardim, O. P. Jensen, T. Kristiansen, C. Longo, C. V. Minte-Vera, C. Minto, I. Mosqueira, G. Chato Osio, D. Ovando, E. R. Selig, J. T. Thorson, J. C.

Walsh, and Y. Ye. 2018. Applying a new ensemble approach to estimating stock status of marine fisheries around the world. *Conserv. Lett.* 11:1-9.

Rossi, S. P., S. P. Cox, H. P. Benoît, and D. P. Swain. 2019. Inferring fisheries stock status from competing hypotheses. *Fish. Res.* 216:155-166.

Sainsbury, K. J. 1988. The ecological basis of multispecies fisheries, and management of a demersal fishery in tropical Australia. In Gulland, J.A., ed. *Fish Population Dynamics*, 2nd edn. Wiley (New York), p. 349–82.

Scott, F., E. Jardim, C. P. Millar, and S. Cerviño. 2016. An applied framework for incorporating multiple sources of uncertainty in fisheries stock assessments. *PLoS ONE* 11:e0154922.

Shertzer, K. W., M. H. Prager, and E. H. Williams. 2008. A probability-based approach to setting annual catch levels. *Fish. Bull*. 106:225-232.

Stewart, I. J., and S. J. D. Martell. 2015. Reconciling stock assessment paradigms to better inform fisheries management. *ICES J. Mar. Sci*. 72:2187-2196.

Terceiro, M. 2002. The summer flounder chronicles: science, politics, and litigation, 1975-2000. *Rev. Fish Biol. Fish*. 11:125-168.

Thompson, G. G. 1992. A Bayesian approach to management advice when stock-recruitment parameters are uncertain. *Fish. Bull.* 90:561-573.

Thompson, G. G. 1996. Application of the Kalman Filter to a stochastic differential equation model of population dynamics. *In* D. J. Fletcher, L. Kavalieris, and B. F. J. Manly (eds.), *Statistics in Ecology and Environmental Monitoring 2: Decision making and risk assessment in biology*, p. 181-203. Otago Conference Series 6, University of Otago Press, Dunedin, New Zealand.

Thompson, G. G. 1999. Optimizing harvest control rules in the presence of natural variability and parameter uncertainty. *In* V. R. Restrepo (ed.), *Proceedings of the 5th National NMFS Stock Assessment Workshop: Providing scientific advice to implement the precautionary approach under the Magnuson-Stevens Fishery Conservation and Management Act*, p. 124-145. NOAA Tech. Memo. NMFS-F/SPO-40. National Marine Fisheries Service, NOAA. 1315 East-West Highway, Silver Spring, MD 20910.

U.S. Court of Appeals, District of Columbia Circuit. 2000. *Natural Resources Defense Council., Inc., et al., Appellants, v. William M. Daley, in his official capacity as Secretary of the United States Department of Commerce, et al., Appellees, Pacific Marine Conservation Council and Alaska Marine Conservation Council, Amicus Curiae*. No. 99-5308.

Walters, C. J. 1975. Optimal harvest strategies for salmon in relation to environmental variability and uncertainty in production parameters. *J. Fish. Res. Board Can.* 32:1777-1784.

Wilberg, M. J., and J. R. Bence. 2008. Performance of deviance information criterion model selection in statistical catch-at-age analysis. *Fish. Res*. 93:212-221.

Zhou, S. 2002. Estimating parameters of derived random variables: comparison of the delta and parametric bootstrap methods. *Trans. Am. Fish. Soc*. 131:667-675.

Tables

Table 1. Cross-conditional relative risks. Color scales extend from red (low) to green (high), with separate scales for each table half: 1) the main body, 2) the "Sum" row, and 3) the "Sum" column.

ra = 0

	Candidate								
Pivot	1	2	3	4	5	6	7	8	Sum
1	0.019	0.032	0.022	0.037	0.017	0.031	0.030	0.018	0.207
2	0.019	0.022	0.030	0.011	0.021	0.013	0.023	0.022	0.161
3	0.008	0.026	0.003	0.021	0.003	0.019	0.016	0.004	0.101
4	0.016	0.018	0.027	0.015	0.018	0.015	0.024	0.019	0.153
5	0.009	0.020	0.002	0.014	0.001	0.015	0.014	0.003	0.077
6	0.018	0.020	0.032	0.016	0.024	0.018	0.029	0.026	0.183
7	0.005	0.011	0.006	0.010	0.003	0.009	0.015	0.005	0.065
8	0.004	0.015	0.001	0.013	0.001	0.011	0.007	0.001	0.054
Sum	0.099	0.164	0.123	0.138	0.087	0.132	0.159	0.099	1.000

ra = 2

	Candidate								
Pivot	1	2	3	4	5	6	7	8	Sum
1	0.000	0.000	0.055	0.000	0.009	0.000	0.000	0.081	0.144
2	0.000	0.000	0.048	0.000	0.000	0.000	0.000	0.000	0.048
3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.095	0.000	0.000	0.000	0.312	0.000	0.407
5	0.000	0.000	0.070	0.000	0.000	0.000	0.000	0.000	0.070
6	0.000	0.000	0.080	0.000	0.097	0.000	0.069	0.084	0.331
7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Sum	0.000	0.000	0.348	0.000	0.106	0.000	0.382	0.165	1.000

Table 2. Cross-conditional correlations between models. The pivot model in each sub-table is highlighted.

1	1	2	3	4	5	6	7	8
1	1.000	0.839	0.153	0.259	0.221	0.548	-0.008	0.177
2	0.839	1.000	0.150	0.210	0.291	0.631	-0.020	0.236
3	0.153	0.150	1.000	0.736	0.927	0.633	0.800	0.820
4	0.259	0.210	0.736	1.000	0.696	0.675	0.693	0.626
5	0.221	0.291	0.927	0.696	1.000	0.720	0.702	0.860
6	0.548	0.631	0.633	0.675	0.720	1.000	0.439	0.641
7	-0.008	-0.020	0.800	0.693	0.702	0.439	1.000	0.798
8	0.177	0.236	0.820	0.626	0.860	0.641	0.798	1.000

2	1	2	3	4	5	6	7	8
1	1.000	0.824	0.278	0.312	0.446	0.582	-0.007	0.255
2	0.824	1.000	0.131	0.218	0.486	0.677	-0.079	0.274
3	0.278	0.131	1.000	0.594	0.719	0.378	0.542	0.476
4	0.312	0.218	0.594	1.000	0.551	0.724	0.635	0.611
5	0.446	0.486	0.719	0.551	1.000	0.610	0.381	0.680
6	0.582	0.677	0.378	0.724	0.610	1.000	0.306	0.582
7	-0.007	-0.079	0.542	0.635	0.381	0.306	1.000	0.729
8	0.255	0.274	0.476	0.611	0.680	0.582	0.729	1.000

3	1	2	3	4	5	6	7	8
1	1.000	0.900	0.047	0.156	0.212	0.729	-0.191	0.106
2	0.900	1.000	-0.060	0.082	0.202	0.835	-0.224	0.114
3	0.047	-0.060	1.000	0.686	0.893	0.207	0.627	0.786
4	0.156	0.082	0.686	1.000	0.671	0.364	0.566	0.659
5	0.212	0.202	0.893	0.671	1.000	0.471	0.463	0.873
6	0.729	0.835	0.207	0.364	0.471	1.000	0.069	0.456
7	-0.191	-0.224	0.627	0.566	0.463	0.069	1.000	0.648
8	0.106	0.114	0.786	0.659	0.873	0.456	0.648	1.000

4	1	2	3	4	5	6	7	8
1	1.000	0.854	0.114	0.239	0.218	0.558	-0.120	0.108
2	0.854	1.000	0.008	0.120	0.224	0.655	-0.183	0.104
3	0.114	0.008	1.000	0.629	0.860	0.424	0.714	0.778
4	0.239	0.120	0.629	1.000	0.597	0.634	0.619	0.680
5	0.218	0.224	0.860	0.597	1.000	0.579	0.537	0.818
6	0.558	0.655	0.424	0.634	0.579	1.000	0.239	0.534
7	-0.120	-0.183	0.714	0.619	0.537	0.239	1.000	0.686
8	0.108	0.104	0.778	0.680	0.818	0.534	0.686	1.000

5	1	2	3	4	5	6	7	8
1	1.000	0.930	0.038	0.298	0.208	0.797	-0.245	-0.017
2	0.930	1.000	-0.074	0.220	0.276	0.832	-0.289	0.014
3	0.038	-0.074	1.000	0.419	0.569	0.116	0.553	0.500
4	0.298	0.220	0.419	1.000	0.484	0.605	0.547	0.640
5	0.208	0.276	0.569	0.484	1.000	0.475	0.267	0.520
6	0.797	0.832	0.116	0.605	0.475	1.000	0.026	0.379
7	-0.245	-0.289	0.553	0.547	0.267	0.026	1.000	0.645
8	-0.017	0.014	0.500	0.640	0.520	0.379	0.645	1.000

6	1	2	3	4	5	6	7	8
1	1.000	0.787	-0.003	0.175	0.059	0.557	-0.139	0.025
2	0.787	1.000	-0.070	0.085	0.161	0.636	-0.179	0.095
3	-0.003	-0.070	1.000	0.723	0.897	0.447	0.769	0.801
4	0.175	0.085	0.723	1.000	0.686	0.639	0.693	0.735
5	0.059	0.161	0.897	0.686	1.000	0.567	0.641	0.889
6	0.557	0.636	0.447	0.639	0.567	1.000	0.331	0.600
7	-0.139	-0.179	0.769	0.693	0.641	0.331	1.000	0.743
8	0.025	0.095	0.801	0.735	0.889	0.600	0.743	1.000

7	1	2	3	4	5	6	7	8
1	1.000	0.882	0.056	0.208	0.318	0.741	-0.164	0.025
2	0.882	1.000	-0.102	0.105	0.312	0.760	-0.237	0.016
3	0.056	-0.102	1.000	0.608	0.622	0.164	0.672	0.451
4	0.208	0.105	0.608	1.000	0.535	0.489	0.575	0.444
5	0.318	0.312	0.622	0.535	1.000	0.513	0.350	0.483
6	0.741	0.760	0.164	0.489	0.513	1.000	0.064	0.316
7	-0.164	-0.237	0.672	0.575	0.350	0.064	1.000	0.656
8	0.025	0.016	0.451	0.444	0.483	0.316	0.656	1.000

8	1	2	3	4	5	6	7	8
1	1.000	0.858	0.215	0.245	0.383	0.704	-0.180	0.147
2	0.858	1.000	0.004	0.141	0.356	0.879	-0.232	0.162
3	0.215	0.004	1.000	0.426	0.661	0.144	0.493	0.409
4	0.245	0.141	0.426	1.000	0.502	0.337	0.510	0.463
5	0.383	0.356	0.661	0.502	1.000	0.525	0.321	0.701
6	0.704	0.879	0.144	0.337	0.525	1.000	-0.023	0.423
7	-0.180	-0.232	0.493	0.510	0.321	-0.023	1.000	0.538
8	0.147	0.162	0.409	0.463	0.701	0.423	0.538	1.000

Table 3. Absolute and relative contributions to expected loss (i.e., risk) from each model for the risk-averse (ra=0) and risk-neutral (ra=2) cases, and also for equal and optimal weighting in each case. Color scales extend from red (low) to green (high), with separate scales for each colored column.

	Equal weighting				Optimal weighting			
	Absolute		Relative		Absolute		Relative	
Model	*ra* =0	*ra* =2	*ra* =0	*ra* =2	*ra* =0	*ra* =2	*ra* =0	*ra* =2
1	0.0015	0.0878	0.1663	0.0850	0.0013	0.0785	0.1543	0.0831
2	0.0016	0.1076	0.1786	0.1041	0.0017	0.1193	0.2009	0.1263
3	0.0010	0.2823	0.1110	0.2732	0.0009	0.2313	0.1025	0.2449
4	0.0013	0.1032	0.1489	0.0999	0.0013	0.1198	0.1600	0.1268
5	0.0006	0.1304	0.0736	0.1262	0.0005	0.0936	0.0627	0.0991
6	0.0020	0.1395	0.2276	0.1350	0.0020	0.1701	0.2392	0.1800
7	0.0003	0.0576	0.0383	0.0557	0.0003	0.0402	0.0308	0.0426
8	0.0005	0.1249	0.0557	0.1209	0.0004	0.0919	0.0495	0.0972
Ensemble	0.0088	1.0333	1.0000	1.0000	0.0084	0.9447	1.0000	1.0000

Table 4. Statistics of the OFL distributions in the risk neutral (ra=0) and risk averse (ra=2) cases. The optimal value for each case is shaded. The cumulative probability corresponding to the optimal value, given the probability mass function for that case, is indicated by p.*

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

Table 5a. Results of cross-validation with respect to model weights for the risk-neutral (ra=0) and risk-averse (ra=2) cases. Color shading extends from red (low) to green (high) in each respective column.

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

Table 5b. Results of cross-validation with respect to expected loss (i.e., risk) for the risk-neutral (ra=0) and risk-averse (ra=2) cases.

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

Figures

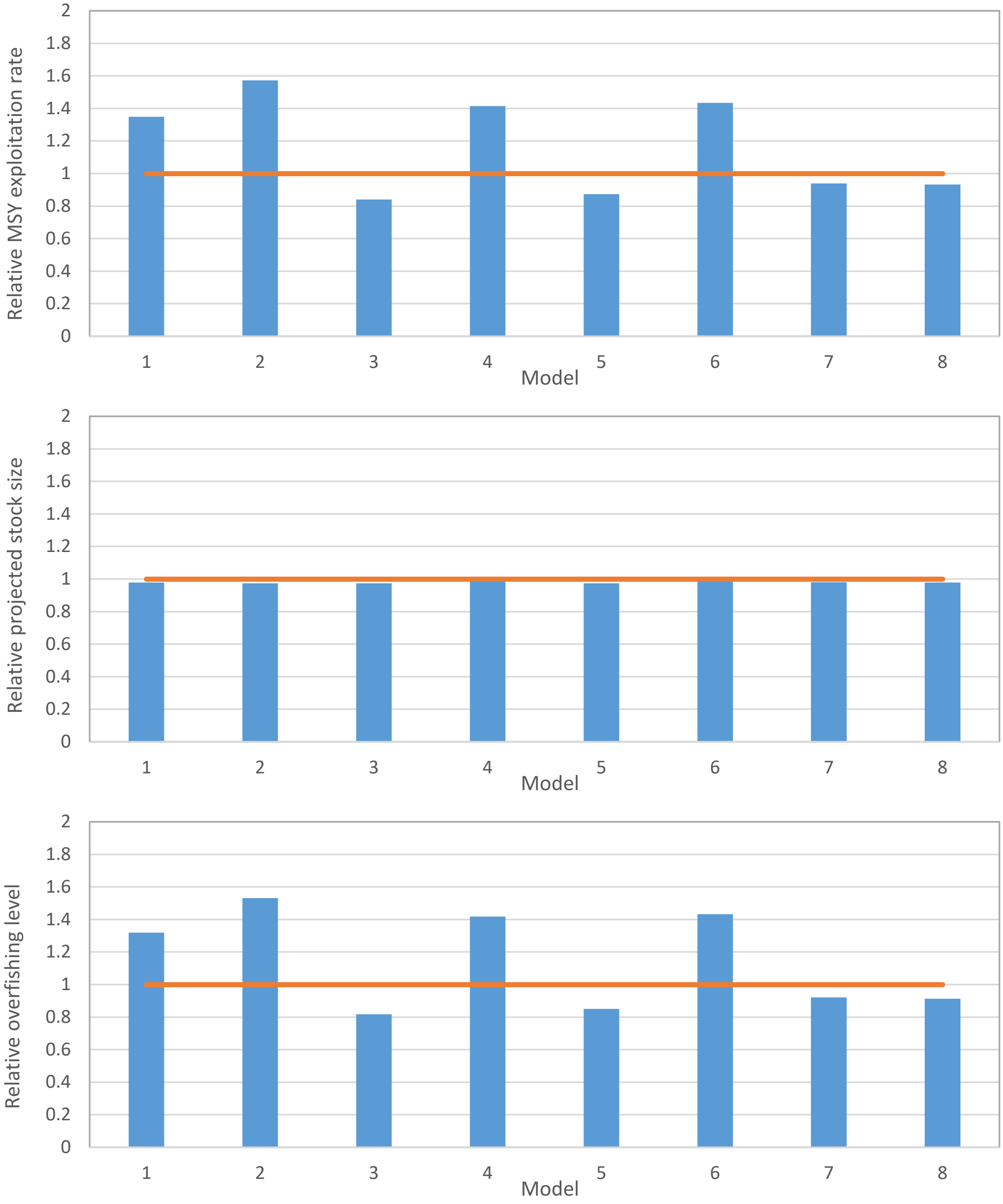

Figure 1. MSY exploitation rate, projected stock size, and overfishing level; all relative to true value.

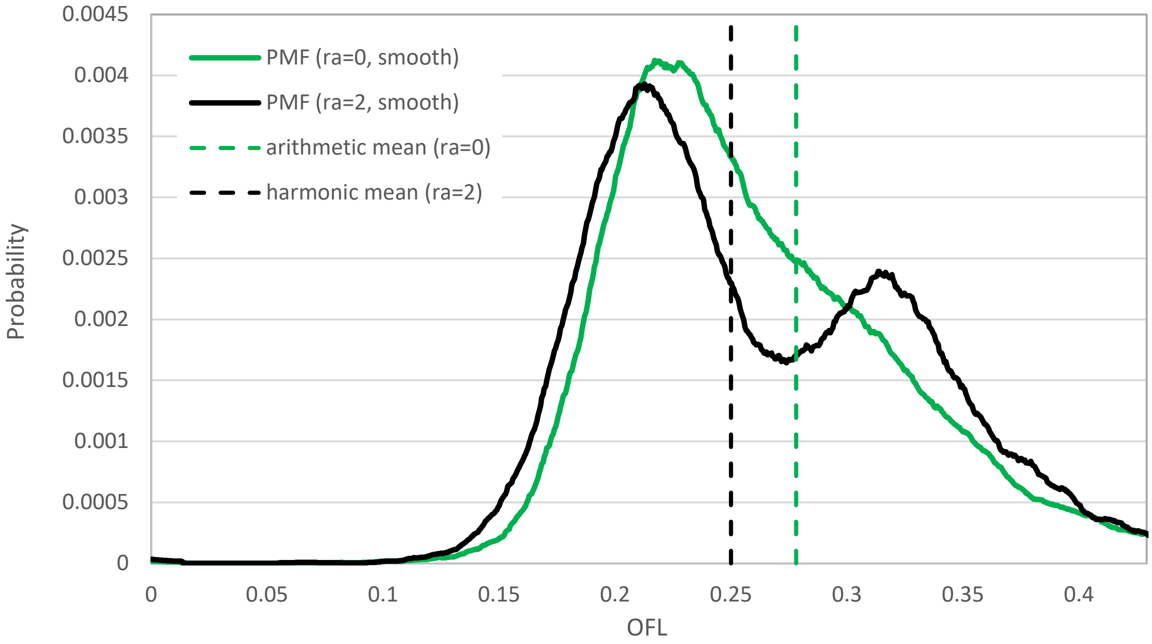

Figure 2a. Probability mass functions under optimzed model weights for the risk-neutral (green) and risk-averse (black) cases. Vertical dashed lines indicate respective optima.

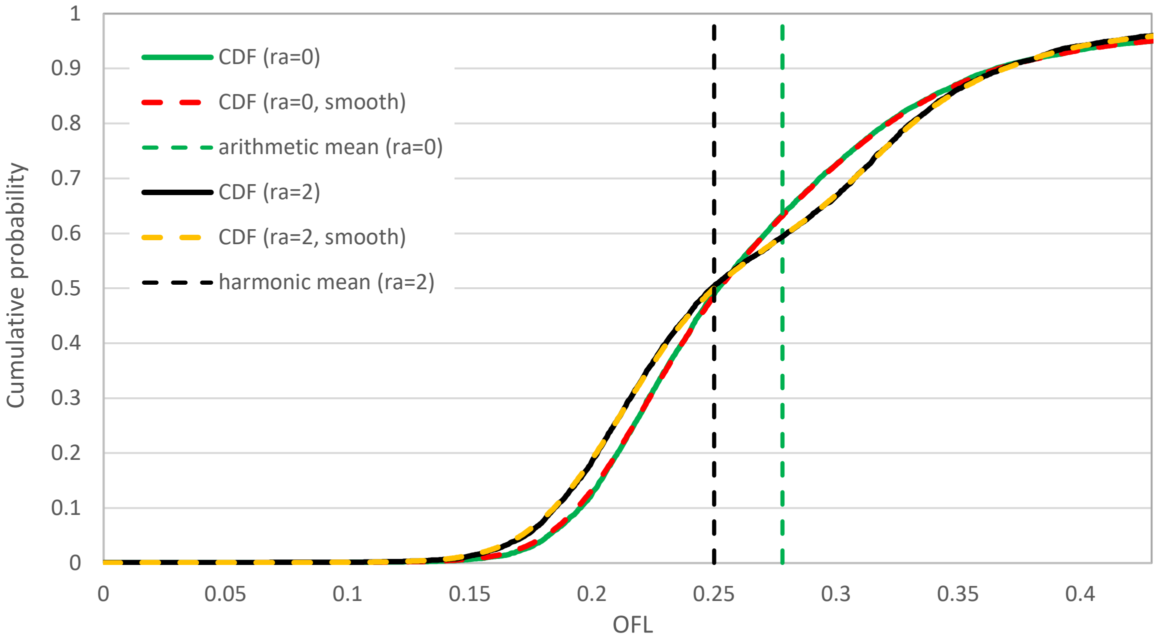

Figure 2b. Cumulative distribution functions under optimized model weights for the risk-neutral (green) and risk-averse (black) cases. Vertical dashed lines indicate respective optima.

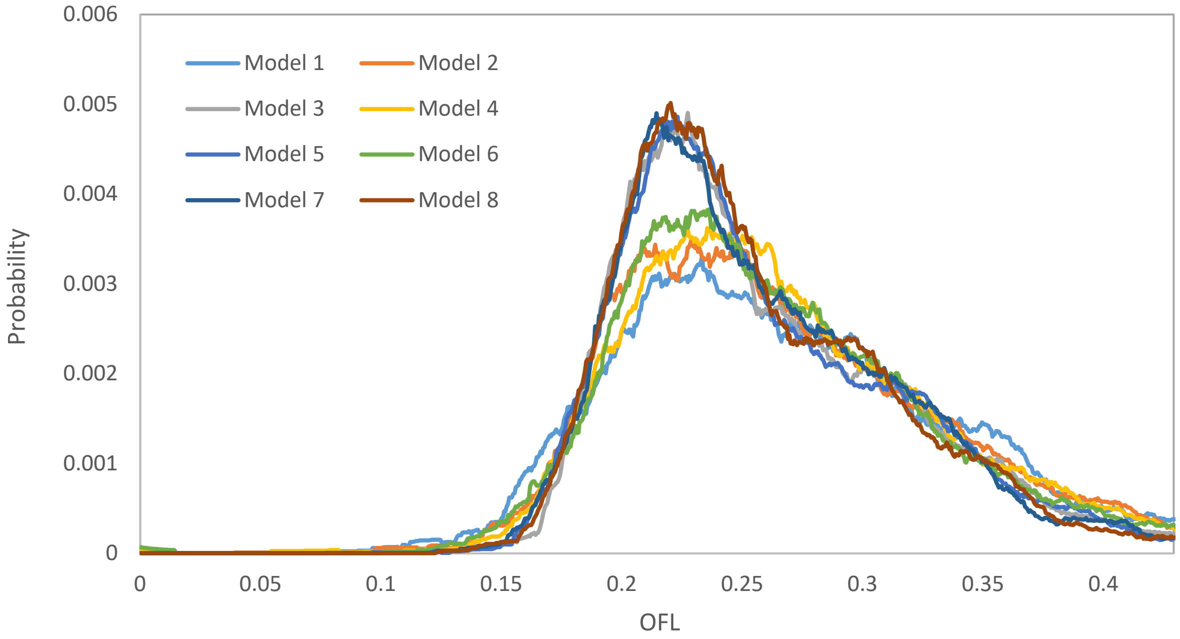

Figure 3a. Smoothed probability mass functions specific to each pivot model (with optimized weights for the risk neutral case, but not multiplied by the respective pivot model's probability).

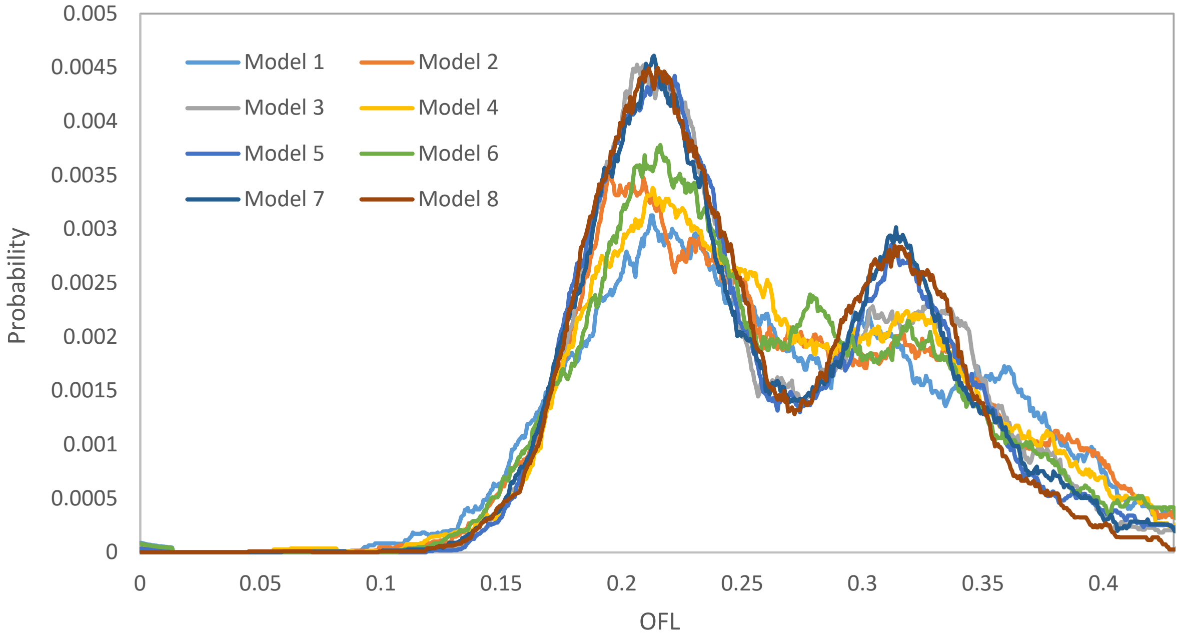

Figure 3b. Smoothed probability mass functions specific to each pivot model (with optimized weights for the risk-averse case, but not multiplied by the respective pivot model's probability).

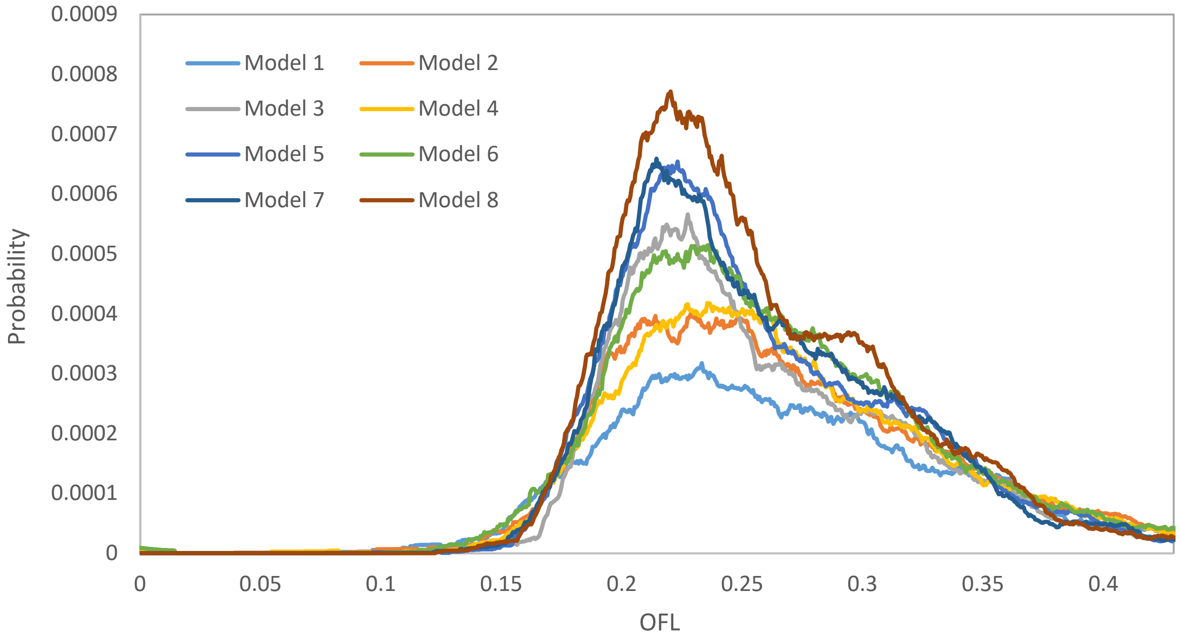

Figure 4a. Smoothed probability mass functions specific to each pivot model (with optimized weights for the risk neutral case, multiplied by the respective pivot model's probability).

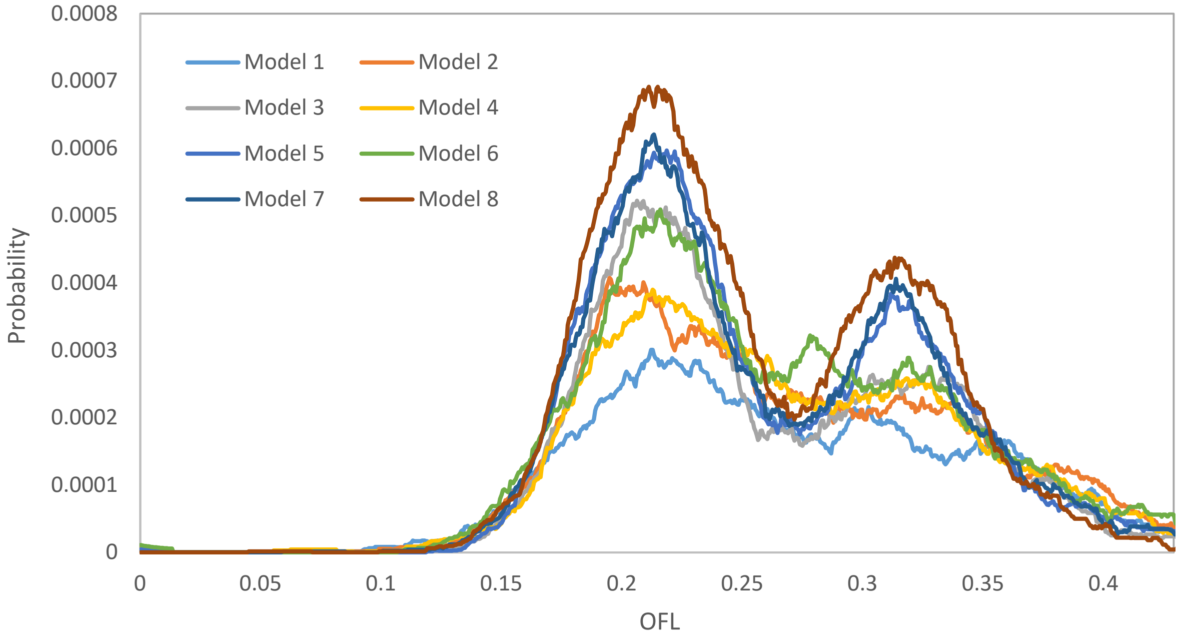

Figure 4b. Smoothed probability mass functions specific to each pivot model (with optimized weights for the risk-averse case, multiplied by the respective pivot model's probability).

Appendix A: Properties of the loss function

Background: constant relative risk aversion

Let relative yield *yrel*(*f*) be defined as a ratio of long-term yields (specifically, yields in the limit as time approaches infinity), where the numerator is the long-term yield obtained under some management-specified fishing mortality rate *f* and the denominator is the maximum sustainable yield (*msy*).

The following loss function formed the basis for the decision-theoretic optima described by Thompson (1992, 1996, 1999) and Restrepo et al. (1998; see section 3.1), as well as the buffer between "acceptable biological catch" and *ofl* in "Tier 1" of the North Pacific Fishery Management Council (NPFMC) groundfish harvest control rules (NPFMC 2018):

$$loss(yrel|f,ra)=\frac{1-yrel(f)^{1-ra}}{1-ra},$$

where *ra* represents risk aversion. In the limit as *ra*→1, the above converges to $-\ln(yrel(f))$.

More precisely, this loss function is said to exhibit "constant relative risk aversion" (CRRA) because it satisfies the following for all values of *yrel* and *ra* (Pratt 1964; Arrow 1965, 1971):

$$ra\equiv -yrel(f)\left(\frac{\frac{d^2loss(yrel|f,ra)}{dyrel(f)^2}}{\frac{dloss(yrel|f,ra)}{dyrel(f)}}\right).$$

Next, let $m_y(q)$ represent the "*y* mean of order *q*" for some positive random variable *y*; that is, the *q*th root of the *q*th noncentral moment (note that *q* need not be an integer; it can be any real number):

$$m_y(q)=\left(\int_0^\infty g_y(y)y^q dy\right)^{1/q},$$

where $g_y(y)$ represents the pdf of *y*, and *k* is an arbitrary constant. In the limit as *q* approaches zero, $m_y(q)$ converges to

$$m_y(0)=exp\left(\int_{-\infty}^\infty g_y(y)\ln(y)dy\right).$$

Familiar examples of order means include the arithmetic (*q* = 1), geometric (*q* = 0), and harmonic (*q* = –1) means. It may be noted that $m_y(q+\Delta)>m_y(q)$ for any positive value of Δ (e.g., Mitrinović 1970).

The risk (i.e., expected loss) in this case is given by

$$risk(f)=\int_0^\infty g_{yrel}(yrel(f))\,loss(yrel|f,ra)dyrel=\frac{1-m_{yrel}(1-ra|f)^{1-ra}}{1-ra}$$

Thus, risk is *minimized* by fishing at the value of *f* that *maximizes* the *yrel* mean of order 1–*ra*.

In Thompson (1992), *ra* was set at unity, so the objective was to maximize the *geometric* mean of relative yield, which, in the particular model considered in that paper, was achieved by fishing at the *harmonic* mean of the *msy* fishing mortality rate.

Loss function used in this analysis

The CRRA loss function is useful for problems like the one described above, where, in the absence of uncertainty, the quantity of interest *itself* is the thing to be minimized ($-yrel(f)$, in the above). Thus, for a very simple harvest control rule in which the fishing mortality rate does not vary, the analysis described above can be used to define that rate. Thus, following that analysis, the *f* that should define the control rule is that which maximizes the *yrel* mean of order 1–*ra*. For example, by setting *ra*=0, maximizing the geometric mean of *yrel* gives a risk-neutral estimate of the *msy* fishing mortality rate.

However, the fact that this is simply an *estimate* of the *msy* fishing mortality rate is significant. Moreover, so are the values of any other parameters that are required to specify a catch limit, such as *ofl*, for some future harvest season (i.e., they are all estimates). Use of *ofl* as the quantity of interest in the CRRA loss function would not be appropriate, because, in the absence of uncertainty, *ofl* is not the thing to be minimized; rather, the *error in estimating it* is the thing to be minimized (obviously, in the absence of uncertainty, minimizing estimation error is a trivial problem, but the concept still applies). The difference is that the CRRA loss function is appropriate when the quantity of interest varies monotonically (and in the "right" direction), but is not appropriate for estimating a state of nature, such as the true-but-unknown value of *ofl*.

An alternative to the CRRA loss function that is appropriate for estimating a state of nature is

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

where *y* is the uncertain state of nature to be estimated and $\hat{y}$ is the estimate (note that *y* must be constrained to non-negative values only). In the limit as *ra* approaches unity, this loss function converges to $(\ln(y)-\ln(\hat{y}))^2$.

Note that this function *does not* exhibit constant relative risk aversion, meaning that *ra* no longer has its original interpretation, although it can be viewed as an *ad hoc* measure of risk aversion.

The risk (i.e., expected loss) in this case is given by

$$risk(\hat{y})=\int_0^\infty g_y(y)\,loss(y|\hat{y},ra)dy=\frac{m_y(2(1-ra))^{2(1-ra)}-2m_y(1-ra)^{1-ra}\hat{y}^{1-ra}+\hat{y}^{2(1-ra)}}{(1-ra)^2}.$$

The derivative of risk w.r.t. $\hat{y}$ is

$$\frac{drisk}{d\hat{y}}=2\hat{y}^{-ra}\left(\frac{\hat{y}^{1-ra}-m_y(1-ra)^{1-ra}}{1-ra}\right),$$

which is solved by setting $\hat{y}=m_y(1-ra)$.

Thus, risk minimization involves a mean of order 1–*ra* is in both the CRRA loss function and the loss function used here. In the former, risk is minimized by fishing at the rate that maximizes the *yrel* mean of order 1−*ra*, whereas in the loss function used here, risk is minimized by setting $\hat{y}$ equal to the *y* mean of order 1−*ra*.

Example behaviors of the loss function for integer values of *ra* ranging from −0.2 to 0.2 are shown, conditional on *y*=10, in Figure A1. The way that *ra* determines the shape of the loss function can be seen by comparing the heights of a particular curve at $y = \hat{y} + \Delta$ and at $y = \hat{y} - \Delta$. The point $y = \hat{y} + \Delta$ corresponds to an *under*estimate, because it implies $\hat{y} = y - \Delta$, whereas the point $y = \hat{y} - \Delta$ corresponds to an *over*estimate, because it implies $\hat{y} = y + \Delta$. Curves with negative values of *ra* are higher at $y = \hat{y} + \Delta$ than at $y = \hat{y} - \Delta$, implying that an underestimate is associated with greater loss than an overestimate of the same magnitude. Conversely, curves with positive values of *ra* are higher at $y = \hat{y} - \Delta$ than at $y = \hat{y} + \Delta$, implying that an overestimate is associated with greater loss than an underestimate of the same magnitude.

A more explicitly quantitative interpretation of the manner in which *ra* “characterizes” the extent to which an underestimate is preferable to an overestimate, can be constructed as follows: First, define an “over-under” ratio in terms of positive and negative displacements Δ from a starting value of unity:

$$overunder(\Delta|\hat{y}, ra) = \frac{loss(\hat{y}(1-\Delta)|\hat{y}, ra)}{loss(\hat{y}(1+\Delta)|\hat{y}, ra)}.$$

Figure A2 illustrates the over-under ratio for the same five values of *ra* used in Figure A1, showing that the curves share a common vertical intercept at a value of 1.0. This figure also shows that the over-under ratio varies considerably with Δ, making it hard to use the over-under ratio, by itself, to give an intuitive meaning to the *ra* parameter.

The *derivative* of the over-under ratio, however, allows for a simple interpretation of *ra*, because the derivative always approaches a lower limit of 2*ra* as Δ approaches 0, as shown in Figure A3.

Decision-theoretic estimates of risk are invariant under positive linear transforms of the loss function. However, because the loss function here is scaled so that the minimum possible loss (achieved at *ratio*=1.0) is zero, ratios of risks are meaningful.

References

Arrow, K.J. 1965. *Aspects of the theory of risk-bearing*. Helsinki: Yrjö Jahnssonin Säätiö. 61 p.

Arrow, K.J. 1971. *Essays in the theory of risk-bearing*. Chicago: Markham. 278 p.

Pratt, J.W. 1964. Risk aversion in the small and in the large. *Econometrica* 32:122-36.

Mitrinović, D. S. 1970. *Analytic inequalities*. Springer-Verlag, Berlin. 400 p.

NPFMC (North Pacific Fishery Management Council). 2018. *Fishery Management Plan for Groundfish of the Bering Sea and Aleutian Islands Management Area*. 605 W. 4th Avenue, Suite 306, Anchorage, AK 99501. 174 p.

Restrepo, V. R., G. G. Thompson, P. M. Mace, W. L. Gabriel, L. L. Low, A. D. MacCall, R. D. Methot, J. E. Powers, B. L. Taylor, P. R. Wade, and J. F. Witzig. 1998. *Technical guidance on the use of precautionary approaches to implementing National Standard 1 of the Magnuson-Stevens Fishery Conservation and Management Act*. NOAA Tech. Memo. NMFS-F/SPO-31. National Marine Fisheries Service, NOAA. 1315 East-West Highway, Silver Spring, MD 20910. 54 p.

Thompson, G. G. 1992. A Bayesian approach to management advice when stock-recruitment parameters are uncertain. *Fish. Bull.* 90:561-573.

Thompson, G. G. 1996. Application of the Kalman Filter to a stochastic differential equation model of population dynamics. *In* D. J. Fletcher, L. Kavalieris, and B. F. J. Manly (eds.), *Statistics in Ecology and Environmental Monitoring 2: Decision making and risk assessment in biology*, p. 181-203. Otago Conference Series 6, University of Otago Press, Dunedin, New Zealand.

Thompson, G. G. 1999. Optimizing harvest control rules in the presence of natural variability and parameter uncertainty. *In* V. R. Restrepo (ed.), *Proceedings of the 5th National NMFS Stock Assessment Workshop: Providing scientific advice to implement the precautionary approach under the Magnuson-Stevens Fishery Conservation and Management Act*, p. 124-145. NOAA Tech. Memo. NMFS-F/SPO-40. National Marine Fisheries Service, NOAA. 1315 East-West Highway, Silver Spring, MD 20910.

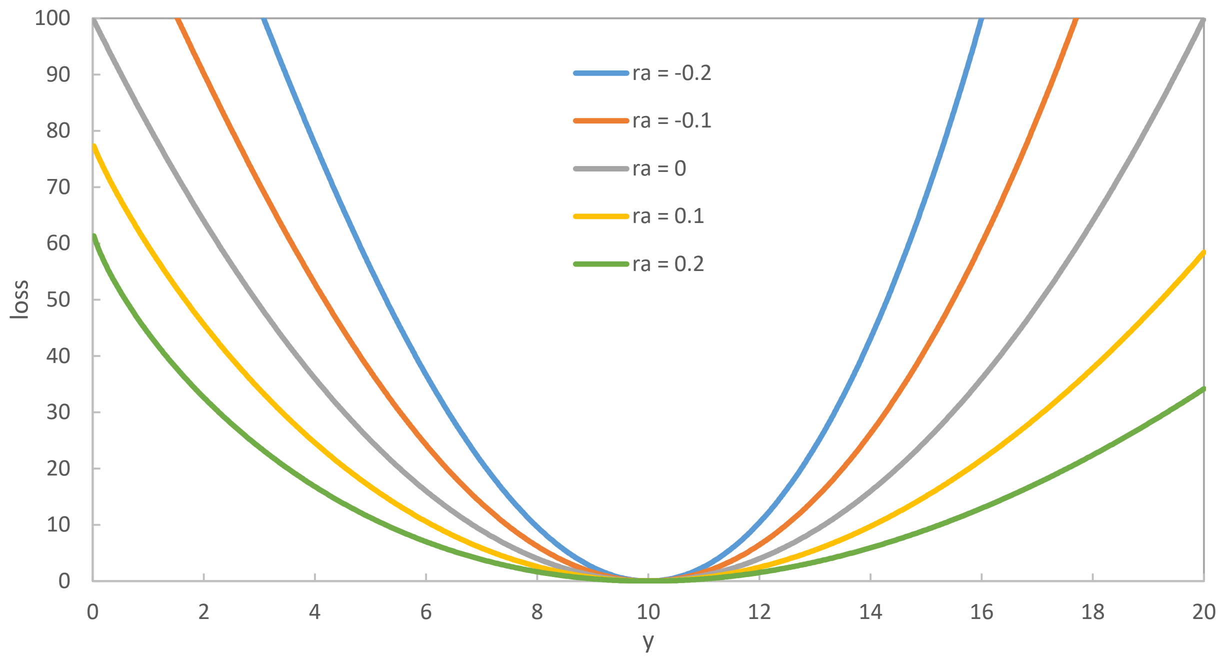

Figure A1. Loss function under five levels of risk aversion, for $\hat{y} = 10$. *A y value less than 10 means that* $\hat{y}$ *is an overestimate, while a y value greater than 10 means that* $\hat{y}$ *is an underestimate.*

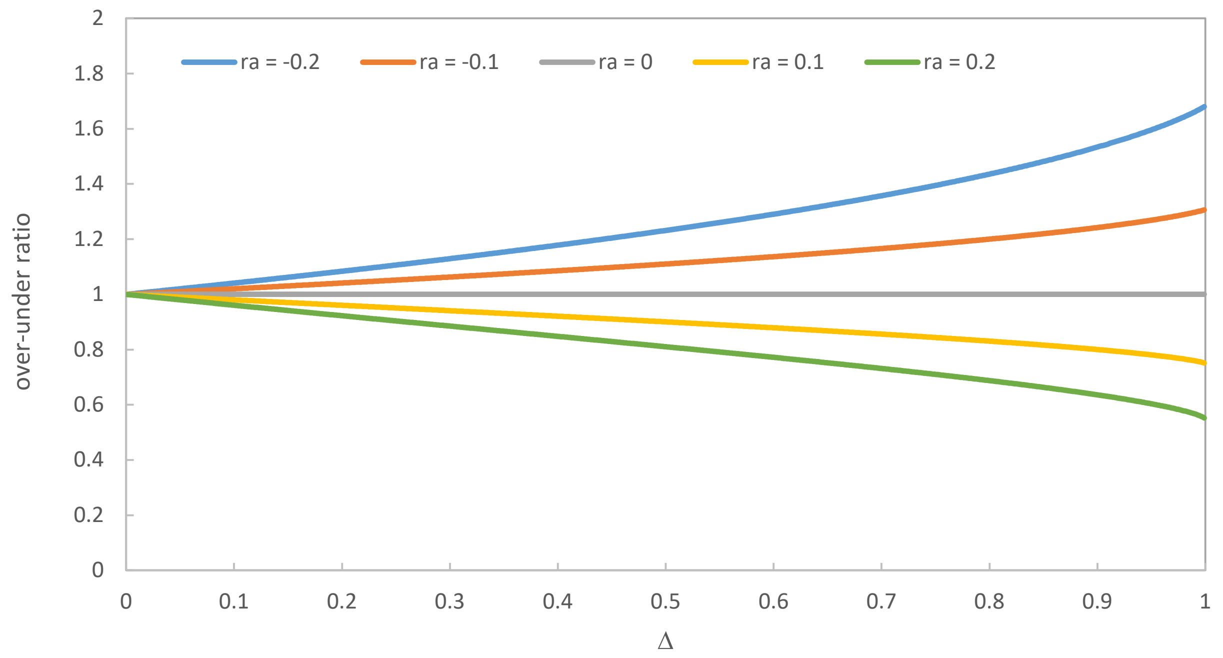

Figure A2a. Ratio of two loss function values, where $\hat{y} = 10$ and y in the numerator and denominator is replaced by $\hat{y}(1 - \Delta)$ and $\hat{y}(1 - \Delta)$, respectively.

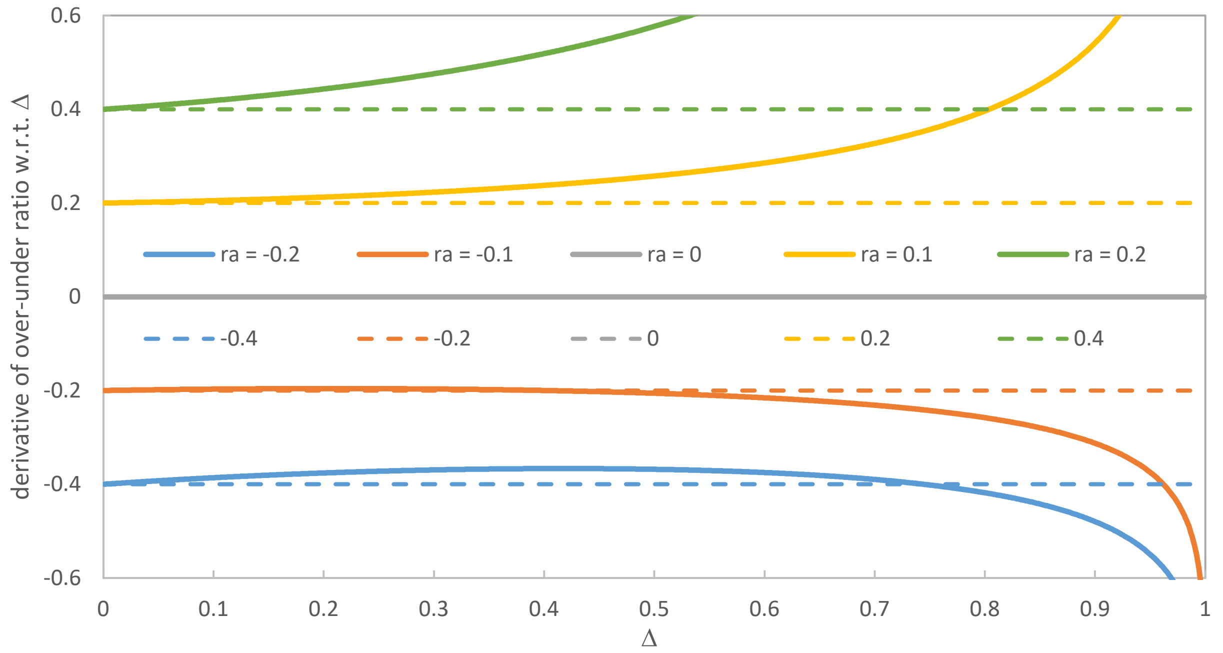

Figure A2b. Derivative of the over-under ratio with respect to Δ, showing limit at 2ra as Δ→0.

Appendix B: Simple Beverton-Holt surplus production model

Structure

An example of a simple assessment model can be based around the following assumptions:

1. Observations of stock size **x** and catch **c** are available for *nt* years.
2. Observed stock size **xobs** is measured, with error $\boldsymbol{\varepsilon}\sim lognormal\left(-\frac{s^2}{2},s^2\right)$, in numbers of fish.
3. Catch is measured, without error, in numbers of fish.
4. Process error is entirely absent.
5. Recruitment in year *t*+1 follows a Beverton-Holt function of the number of fish in year *t*.
6. Fishing mortality is described either as a constant discrete annual rate $\bar{u}$ or as a vector of time-varying discrete annual rates **u**.
7. Natural mortality is described either as a constant discrete annual rate $\bar{v}$ or as a vector of time-varying discrete annual rates $\mathbf{v}=\bar{v}+\mathbf{Zd}$, where the columns of **Z** represent *nvar* time-varying environmental variables measured without error and where **d** is a vector of coefficients.
8. The processes of recruitment, fishing mortality, and natural mortality do not overlap intra-annually, and occur in the following order:
    a. Recruitment
    b. Fishing mortality
    c. Natural mortality

In a simulation/estimation context, an ensemble of such models can be developed easily by dropping the assumption that the identity of **Z** is known, and instead generating a set of number of environmental variables that could *potentiall*y affect natural mortality, identifying each possible subset of those variables, and associating one model with each such subset.

The transition from x_t to x_{t+1} can be written

$$x_{t+1}=x_t+\frac{abx_t}{b+x_t}-\left(u_t+v_t(1-u_t)\right)x_t\,,$$

where *a* is the slope of the stock-recruitment curve at the origin and *b* is the stock size at which a tangent through the origin intersects the asymptotic recruitment level *ab*.

Equilibrium stock size and yield under constant $u_t=\bar{u}$ and $v_t=\bar{v}$ are given by

$$xequ(\bar{u})=\left(\frac{a}{\bar{u}+\bar{v}(1-\bar{u})}-1\right)b\ \text{ and }\ yequ(\bar{u})=\left(\frac{a}{\bar{u}+\bar{v}(1-\bar{u})}-1\right)b\bar{u}\,.$$

The equilibrium fishing mortality rate that sets equilibrium stock size equal to zero is

$$uext=min\left(1,\frac{a-\bar{v}}{1-\bar{v}}\right).$$

Maximum sustainable yield (*msy*) and fishing mortality at *msy* are given by

$$msy=\left(\frac{a+\bar{v}-2\sqrt{a\bar{v}}}{1-\bar{v}}\right)b\ \text{ and }\ umsy=\frac{\sqrt{a\bar{v}}-\bar{v}}{1-\bar{v}}\,.$$

In order to keep *umsy* within the range (0,1), the following constraint is necessary: $\bar{v} < a < 1/\bar{v}$. Note that the popular rule of thumb in which the fishing mortality rate at *msy* equals the natural mortality rate will hold if and only if $a = (2 - \bar{v})^2\bar{v}$.

Simulation

For the simulation that formed the basis of the analysis described in the main text, nt was set at a value of 40, and an ensemble was formed by including three columns in the "master" $\mathbf{Z}$ matrix, then associating a model with each possible subset of those columns (including the null subset), giving a total of eight models, configured as shown below:

Model:	1	2	3	4	5	6	7	8
nvar:	0	1	1	1	2	2	2	3
columns of **Z**:	none	1	2	3	1,2	1,3	2,3	1,2,3

The parameter values used were as follow (note: the scale of the stock size is arbitrary):

Parameter:	a	b	d_i	s	$\bar{v}$	x_1
Value:	0.648	1.000	0.024	0.050	0.200	2.240

(Notes: 1) The *a* parameter was set, conditional on $\bar{v}$, so as to satisfy the *umsy*= $\bar{v}$ rule of thumb. 2) The elements of **d** were set, conditional on $\bar{v}$ the maximum value of *nvar* across all models in the ensemble, at a single positive value such that the chance of achieving a negative natural mortality rate was approximately 1E-06. 3) Initial stock size x_1 was set at the equilibrium unexploited level *xequ*(0), conditional on *a*, *b*, and $\bar{v}$.)

To generate **x**, Model 5 was chosen as the true model.

To generate **u**, each value of u_t (t = 1, ..., nt) was drawn randomly from a lognormal distribution with a median value set at 75% of *umsy* and a log-scale standard deviation set such that the probability of exceeding *umsy* in any single year was 5%.

To generate **xobs**, each value of $xobs_t$ (t = 1, ..., nt) was generated as the product of x_t and an error term ε_t drawn randomly from a mean-unbiased lognormal distribution with log-scale standard deviation s.

To generate **c**, each value of c_t was generated as the product of x_t and u_t (t = 1, ..., nt).

To generate **Z**, each element was drawn randomly from a (0,1) normal distribution, then each column was normalized to exhibit zero mean and unit variance.

The simulated time series of **x** and **u** (true values, unknown to the estimation algorithm) and **xobs**, **c**, and **Z** (data, known to the estimation algorithm) are shown in Table B1.

The time series of **v** for each model, implied by the equation $\mathbf{v} = \bar{v} + \mathbf{Z}(\text{model})\mathbf{d}$ (where $\bar{v}$ and **d** are true values, unknown to the estimation algorithm, and each **Z**(model) consists of the model-specific columns of **Z**, known to the estimation algorithm), are shown in Table B2. Each column in Table B2 has a mean equal to the true value of $\bar{v}$ (0.2), with model-specific minima and maxima as shown below:

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Min:	0.200	0.148	0.146	0.160	0.121	0.131	0.113	0.087
Max:	0.200	0.254	0.249	0.251	0.277	0.269	0.270	0.292

Estimation

Parameters were estimated by the method of maximum likelihood, where most parameters were either log- or logit-transformed in order to remove potential difficulties associated with parameter estimates hitting logical bounds. Specifically:

- Parameters left on the natural scale: **d**
- Log-transformed parameters: b, s, x_1
- Logit-transformed parameters: a, $\bar{v}$

The logit transform of $\bar{v}$ was computed in the usual way, viz., $ln(\bar{v}/(1-\bar{v}))$. However, due to the constraint on feasible values of a described above (i.e., $\bar{v} < a < 1/\bar{v}$), the logit transform of a needed to be conditional on $\bar{v}$, taking the form

$$logit(a) = ln\left(\frac{(a-\bar{v})\bar{v}}{1-a\bar{v}}\right).$$

Note that, given an estimate of each x_t, the corresponding u_t can be estimated in closed form as c_t/x_t (recall that c_t is assumed to be measured without error), so **u** was simply a transform of the data and the statistically estimated parameters. The total number of statistically estimated parameters is only 5+*nvar*.

Table B1. Time series of states (true values, unknown to the estimation algorithm) and data (known to the estimation algorithm) used in the simple Beverton-Holt surplus production model.

	True values		Data				
t	**x**	**u**	**xobs**	**c**	$\mathbf{Z}^{<1>}$	$\mathbf{Z}^{<2>}$	$\mathbf{Z}^{<3>}$
1	2.240	0.166	2.379	0.373	-0.857	-2.207	-1.390
2	2.081	0.135	1.973	0.281	0.099	0.600	-1.566
3	1.847	0.128	1.926	0.237	-0.038	-0.288	0.279
4	1.722	0.155	1.818	0.267	-1.510	0.719	-0.233
5	1.601	0.202	1.554	0.324	0.863	1.374	0.112
6	1.352	0.212	1.331	0.287	-0.354	2.022	-1.054
7	1.181	0.199	1.247	0.235	1.340	-1.125	-0.005
8	1.103	0.126	1.078	0.140	-0.552	0.664	-0.979
9	1.108	0.136	1.125	0.151	-0.754	0.346	0.886
10	1.115	0.156	1.219	0.174	0.379	0.961	0.311
11	1.064	0.144	1.052	0.153	0.599	0.470	0.930
12	1.039	0.168	1.030	0.174	1.842	0.203	0.499
13	0.979	0.183	1.054	0.179	-0.245	-0.407	-0.820
14	0.973	0.162	0.943	0.158	-2.160	-0.755	-0.674
15	1.030	0.133	1.020	0.137	-1.470	-1.786	-1.378
16	1.114	0.154	1.057	0.172	-0.249	-0.613	-1.226
17	1.115	0.155	1.082	0.173	-0.032	-1.720	-1.267
18	1.135	0.139	1.119	0.158	1.183	-0.752	1.141
19	1.116	0.174	1.114	0.194	1.345	1.121	-0.401
20	1.024	0.134	1.033	0.138	-0.810	-1.175	-1.419
21	1.080	0.160	1.176	0.173	-0.763	-0.516	2.110
22	1.090	0.161	1.072	0.175	-1.340	1.643	-0.764
23	1.063	0.164	1.012	0.174	0.478	-0.163	1.466
24	1.038	0.163	0.993	0.169	1.406	-1.250	-0.118
25	1.022	0.175	1.049	0.179	-0.098	-1.277	0.214
26	1.030	0.142	1.058	0.146	0.054	0.299	1.184
27	1.028	0.174	1.013	0.179	0.914	0.688	1.030
28	0.975	0.190	0.986	0.186	0.451	-0.596	1.132
29	0.954	0.115	0.921	0.109	2.235	0.926	0.621
30	0.927	0.131	0.908	0.121	-1.261	0.018	0.853
31	0.981	0.144	1.098	0.141	-0.551	2.036	0.857
32	0.962	0.137	0.958	0.131	-0.694	0.923	0.738
33	0.978	0.120	0.933	0.117	-0.430	-0.129	-0.959
34	1.020	0.148	1.029	0.151	0.487	0.120	1.653
35	1.010	0.151	1.028	0.152	-0.002	0.425	-0.552
36	1.003	0.137	0.898	0.137	0.702	0.055	-1.647
37	1.001	0.171	1.043	0.172	0.456	-0.118	-0.200
38	0.981	0.134	0.947	0.131	-1.593	-0.115	0.454
39	1.036	0.131	1.012	0.136	1.025	0.096	0.584
40	1.025	0.150	0.948	0.154	-0.097	-0.715	-0.400

Table B2. Model-specific time series of natural mortality (true values, conditional on the functional form of the respective model).

Natural mortality **v**							
Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
0.200	0.179	0.146	0.166	0.126	0.145	0.113	0.092
0.200	0.202	0.215	0.162	0.217	0.164	0.177	0.179
0.200	0.199	0.193	0.207	0.192	0.206	0.200	0.199
0.200	0.163	0.217	0.194	0.181	0.158	0.212	0.175
0.200	0.221	0.233	0.203	0.254	0.224	0.236	0.257
0.200	0.191	0.249	0.174	0.241	0.166	0.224	0.215
0.200	0.233	0.173	0.200	0.205	0.232	0.173	0.205
0.200	0.187	0.216	0.176	0.203	0.163	0.192	0.179
0.200	0.182	0.208	0.222	0.190	0.203	0.230	0.212
0.200	0.209	0.223	0.208	0.233	0.217	0.231	0.240
0.200	0.215	0.211	0.223	0.226	0.237	0.234	0.249
0.200	0.245	0.205	0.212	0.250	0.257	0.217	0.262
0.200	0.194	0.190	0.180	0.184	0.174	0.170	0.164
0.200	0.148	0.182	0.184	0.129	0.131	0.165	0.113
0.200	0.164	0.157	0.167	0.121	0.131	0.123	0.087
0.200	0.194	0.185	0.170	0.179	0.164	0.155	0.149
0.200	0.199	0.158	0.169	0.157	0.168	0.127	0.127
0.200	0.229	0.182	0.228	0.210	0.256	0.209	0.238
0.200	0.233	0.227	0.190	0.260	0.223	0.217	0.250
0.200	0.180	0.171	0.166	0.152	0.146	0.137	0.117
0.200	0.181	0.187	0.251	0.169	0.233	0.239	0.220
0.200	0.167	0.240	0.181	0.207	0.149	0.221	0.189
0.200	0.212	0.196	0.236	0.208	0.247	0.232	0.243
0.200	0.234	0.170	0.197	0.204	0.231	0.167	0.201
0.200	0.198	0.169	0.205	0.167	0.203	0.174	0.172
0.200	0.201	0.207	0.229	0.209	0.230	0.236	0.237
0.200	0.222	0.217	0.225	0.239	0.247	0.242	0.264
0.200	0.211	0.186	0.227	0.196	0.238	0.213	0.224
0.200	0.254	0.222	0.215	0.277	0.269	0.238	0.292
0.200	0.169	0.200	0.221	0.170	0.190	0.221	0.191
0.200	0.187	0.249	0.221	0.236	0.207	0.270	0.257
0.200	0.183	0.222	0.218	0.206	0.201	0.240	0.224
0.200	0.190	0.197	0.177	0.186	0.166	0.174	0.163
0.200	0.212	0.203	0.240	0.215	0.252	0.243	0.255
0.200	0.200	0.210	0.187	0.210	0.187	0.197	0.197
0.200	0.217	0.201	0.160	0.218	0.177	0.161	0.178
0.200	0.211	0.197	0.195	0.208	0.206	0.192	0.203
0.200	0.161	0.197	0.211	0.159	0.172	0.208	0.170
0.200	0.225	0.202	0.214	0.227	0.239	0.216	0.241
0.200	0.198	0.183	0.190	0.180	0.188	0.173	0.171

Appendix C: OFL histograms

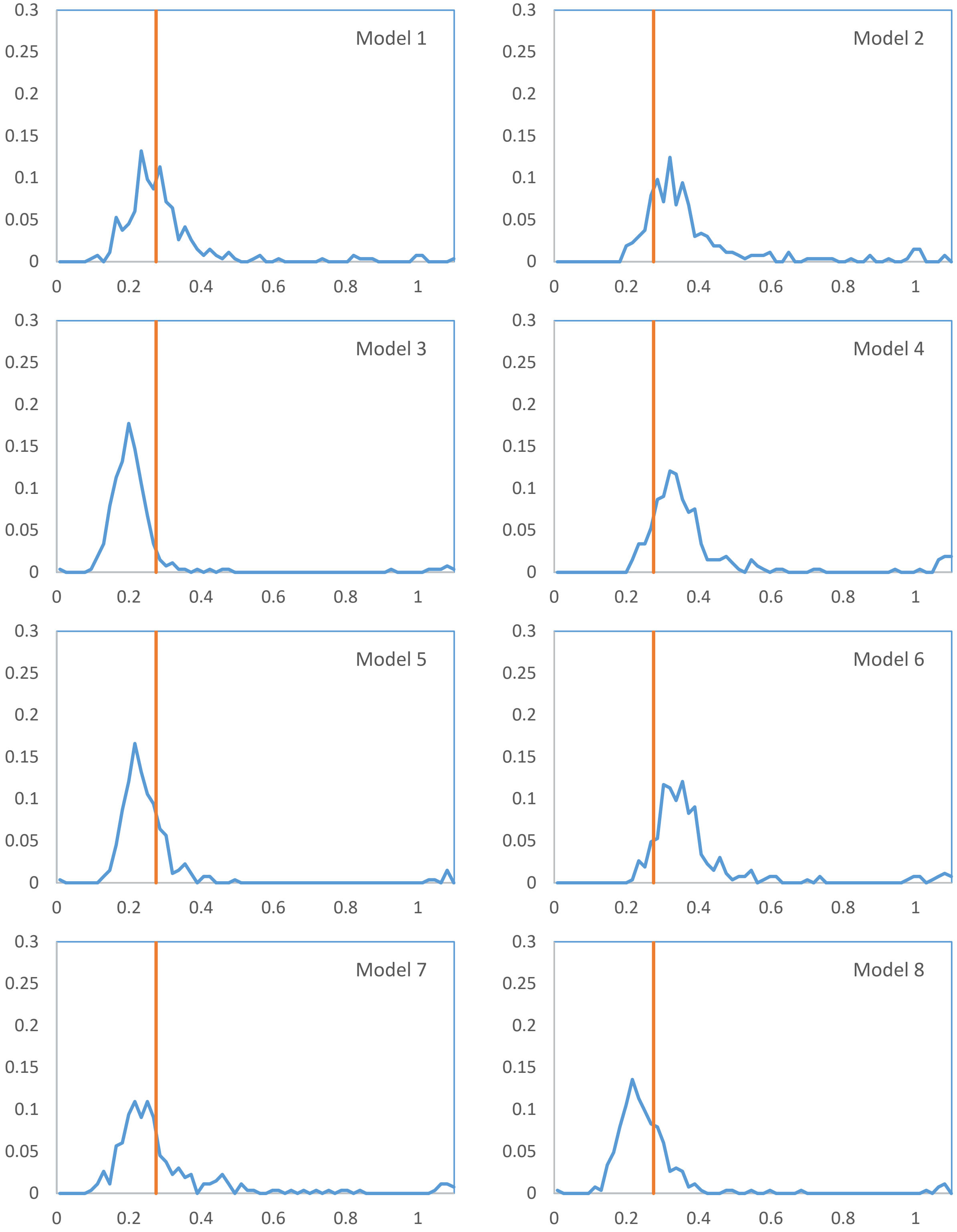

Figure C1. Histogram of OFL for each candidate model, given pivot = Model 1.

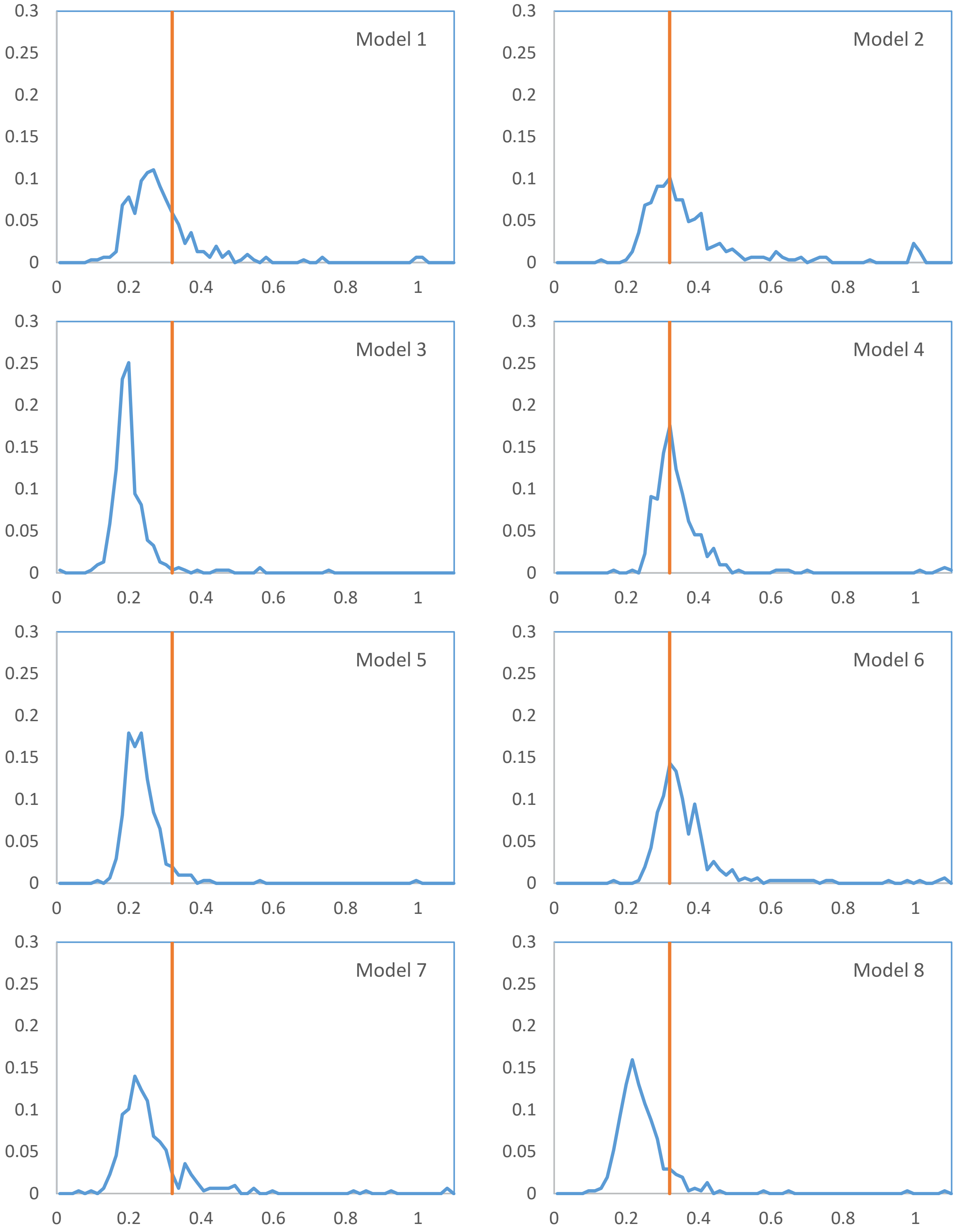

Figure C2. Histogram of OFL for each candidate model, given pivot = Model 2.

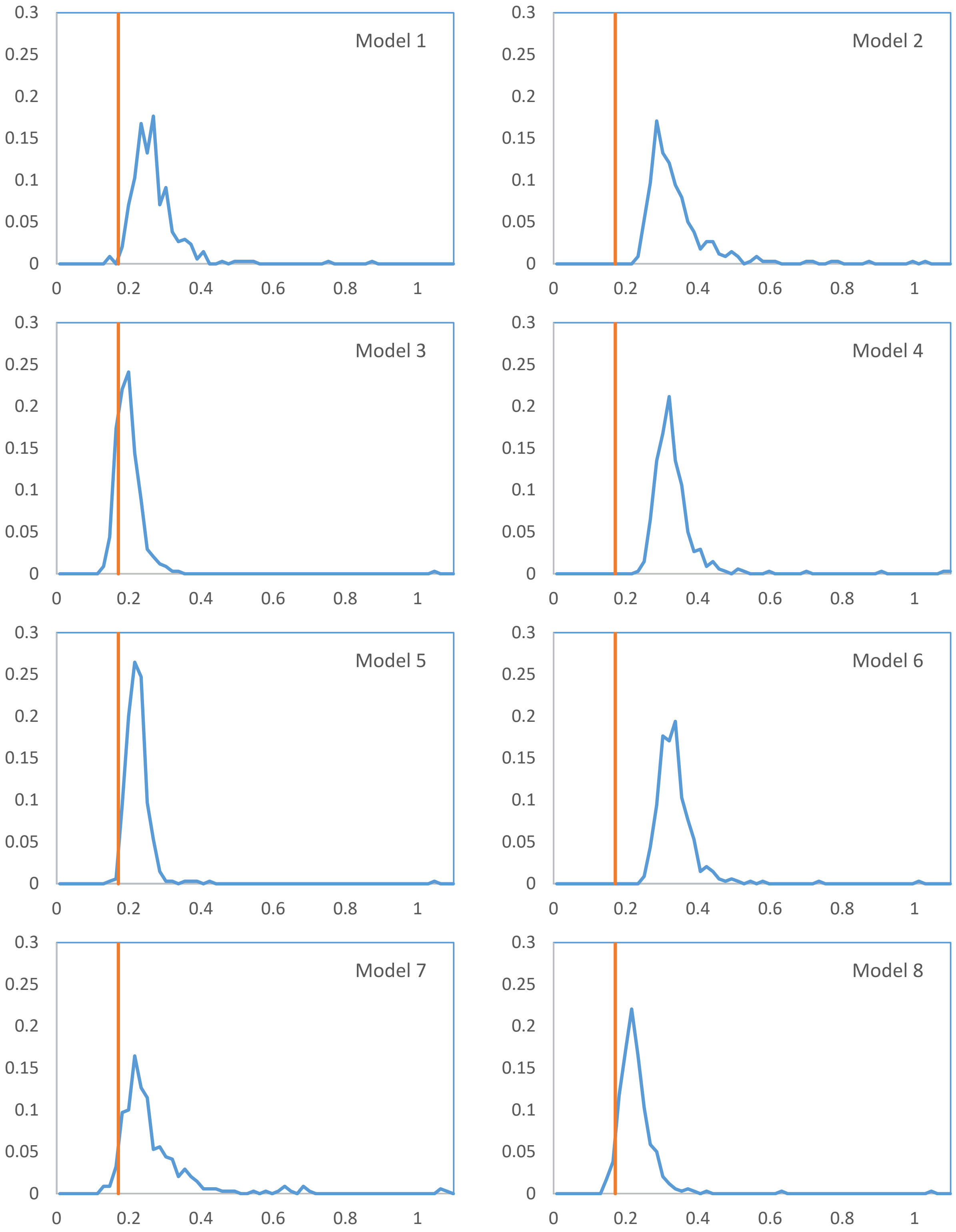

Figure C3. Histogram of OFL for each candidate model, given pivot = Model 3.

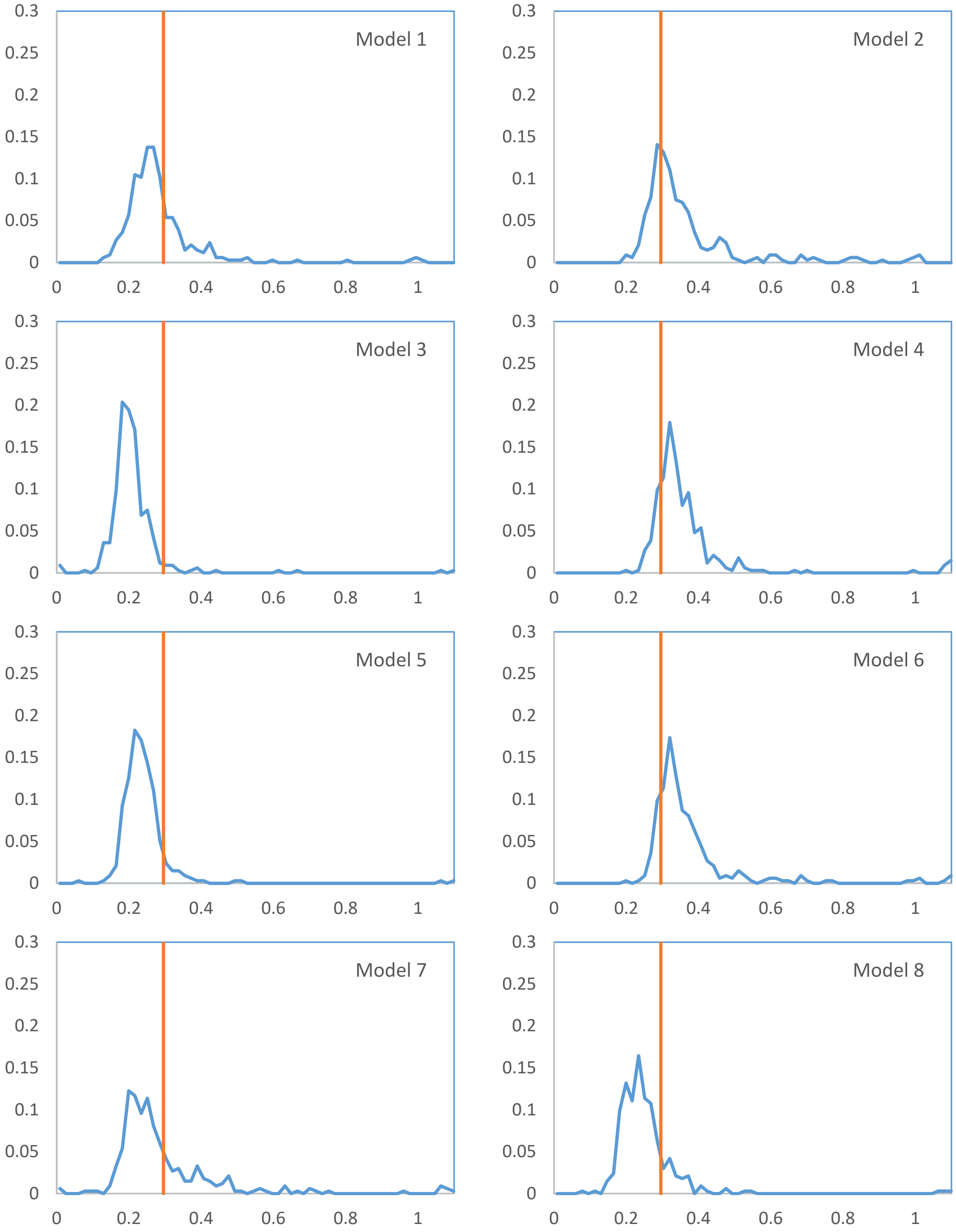

Figure C4. Histogram of OFL for each candidate model, given pivot = Model 4.

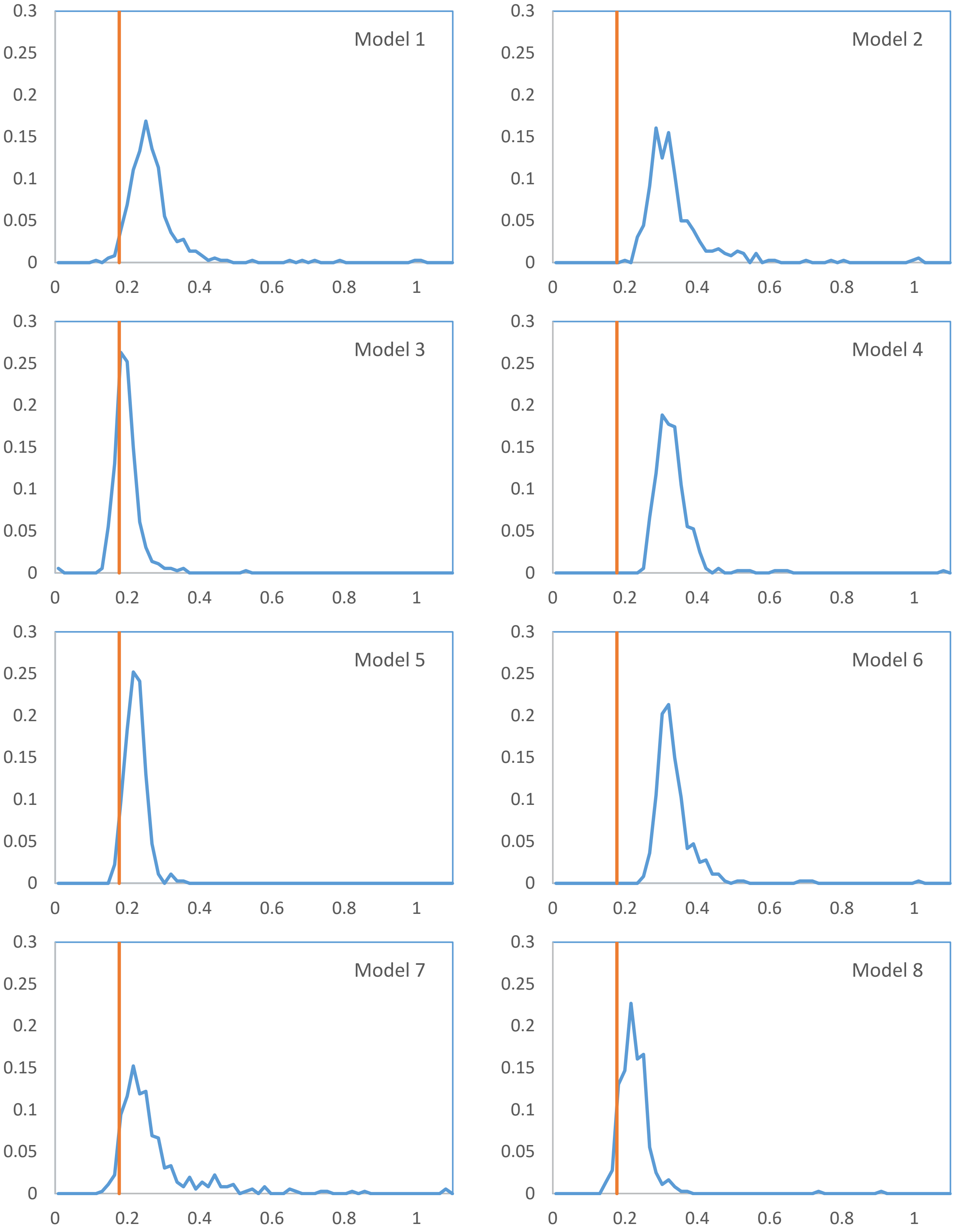

Figure C5. Histogram of OFL for each candidate model, given pivot = Model 5.

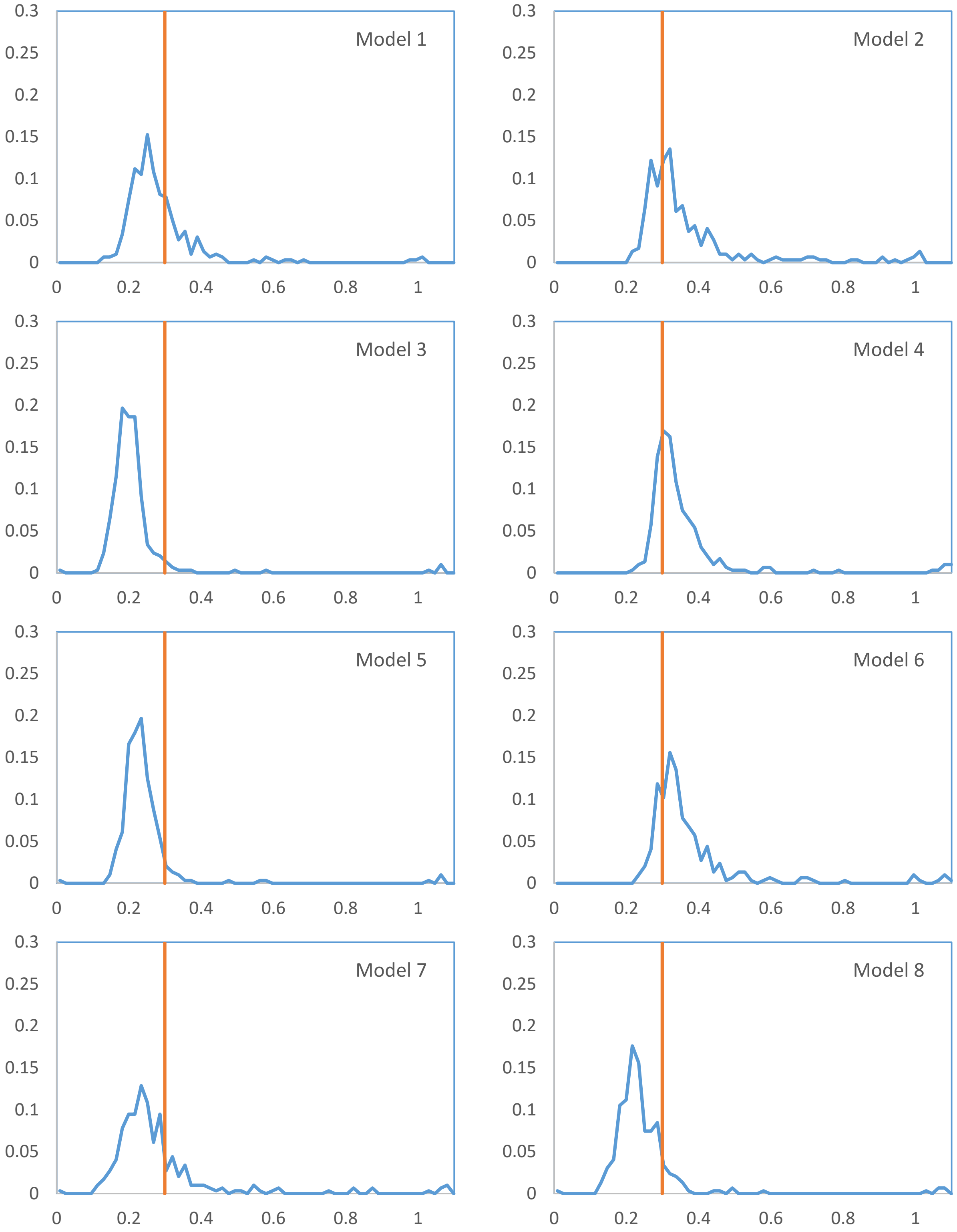

Figure C6. Histogram of OFL for each candidate model, given pivot = Model 6.

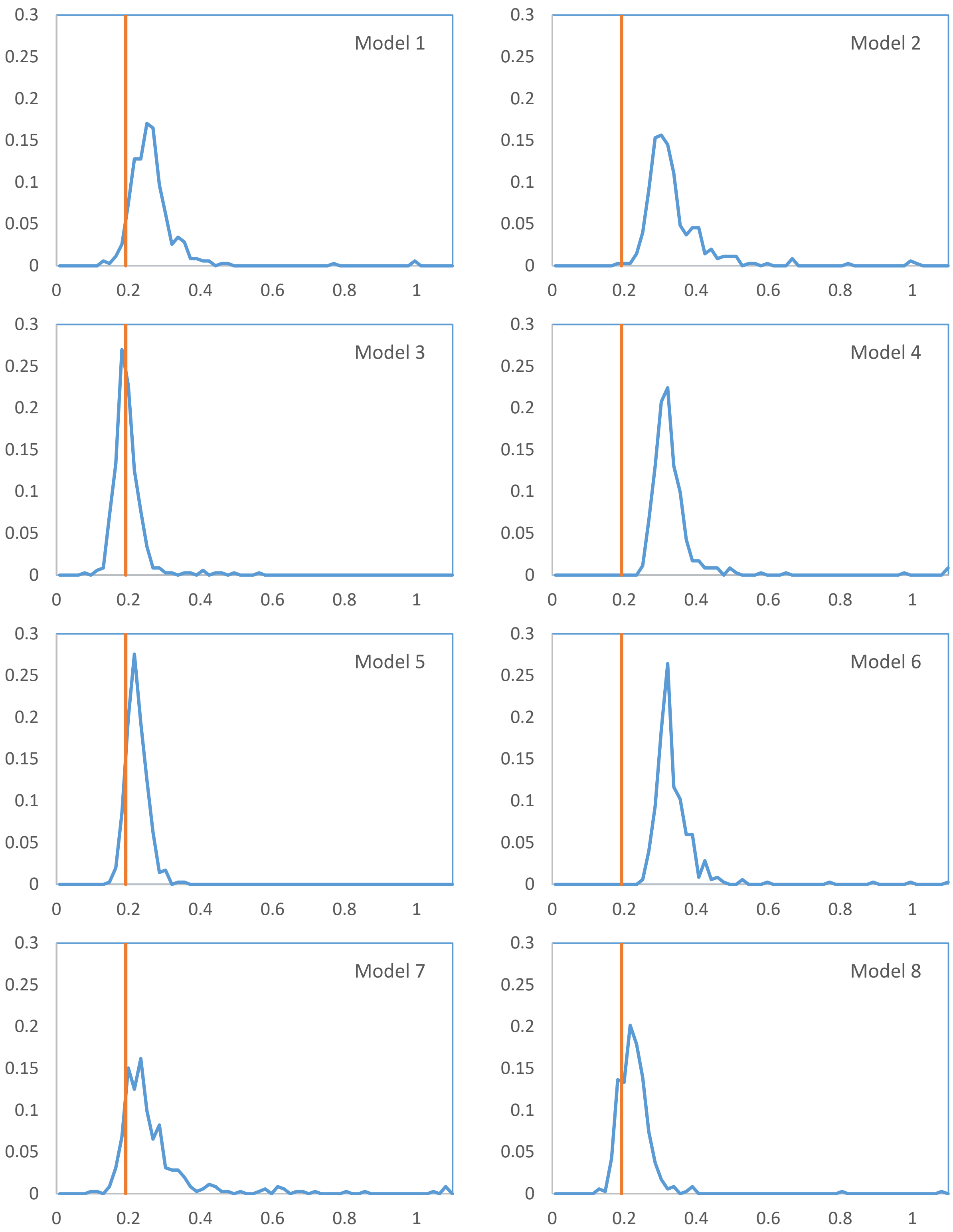

Figure C7. Histogram of OFL for each candidate model, given pivot = Model 7.

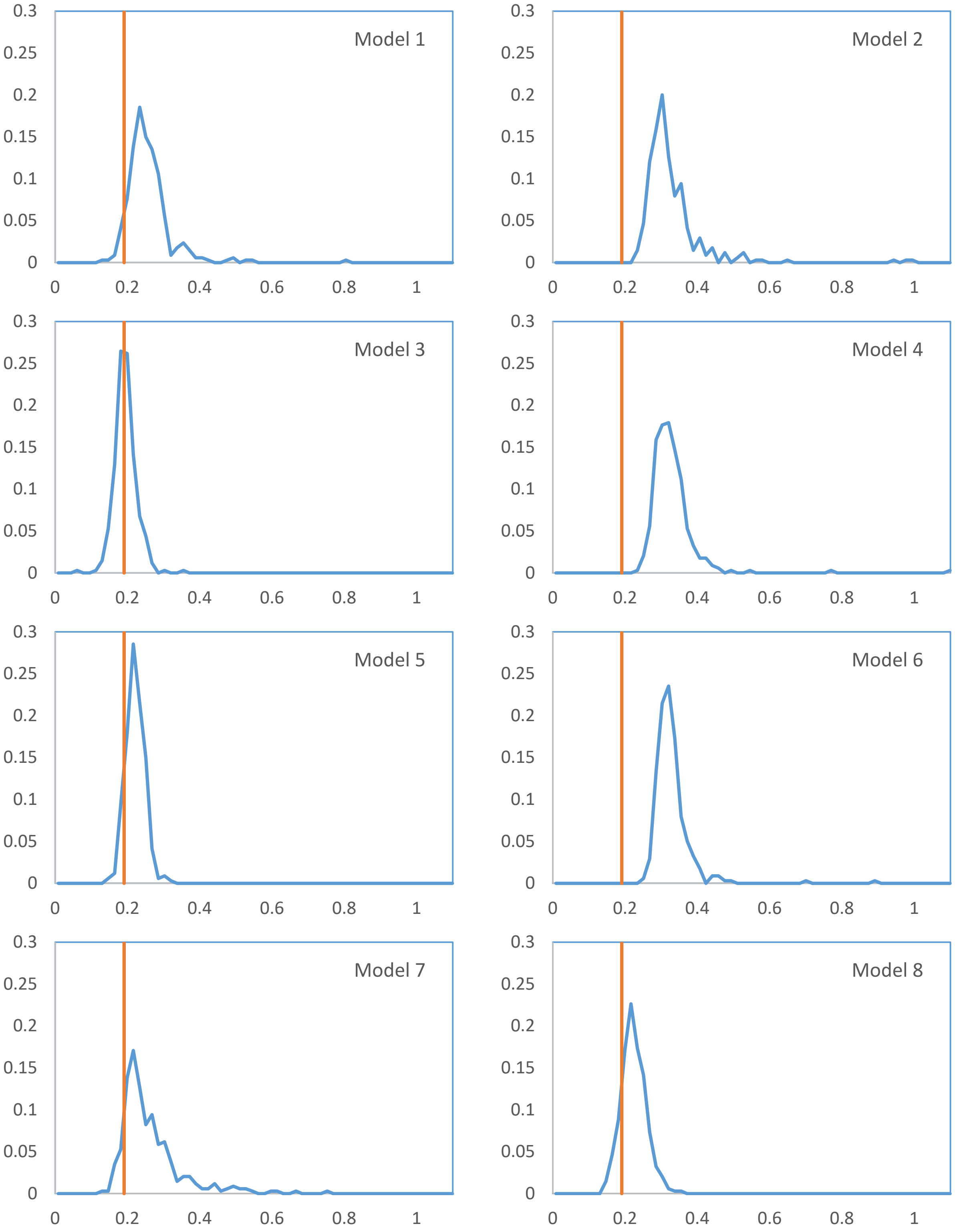

Figure C8. Histogram of OFL for each candidate model, given pivot = Model 8.