Cross-conditional model averaging: A potential tool for improving stock assessment estimates

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 model averaging (CCMA), a method of model averaging that treats each model in a set, one at a time, as though it were the true model, then develops an optimal set of weights for the ensemble, conditional on that assumption, and then finally averages across the set of conditionally optimized ensembles, thus enabling estimation based on an "ensemble of ensembles." A pair of simple examples is provided to illustrate the method: one based on estimation of order means from two candidate statistical distributions, and the other based on estimation of typical population dynamic parameters from two candidate stock assessment models. When performance is measured in terms of expected mean squared error, CCMA can perform no worse than either the best single-model approach or "conventional" model averaging, and the results from the two simple examples demonstrate that CCMA can sometimes perform substantially better than either of these.

Introduction

Providing accurate point estimates of quantities such as stock size, recruitment strengths, and reference fishing mortality rates has long been an emphasis of fishery stock assessment, and the use of statistical assessment models has improved the ability of assessment scientists to do so (e.g., Hilborn 2003). Typically, the assessment scientist develops several statistical models, evaluates them, and then chooses the "best" one, from which point estimates of the relevant quantities are derived. However, assessment scientists have also long recognized (at least as far back as Walters 1975) that provision of accurate point estimates alone is insufficient, and that this should be supplemented with provision of accurate estimates of the *uncertainty* associated with those point estimates. As with the point estimates themselves, the estimates of uncertainty have typically been based on a single "best" model. Recently, though, model averaging (or ensemble modeling; these terms will be viewed as synonymous here) has been advocated in the fishery science literature as providing both more accurate point estimates and more accurate estimates of uncertainty than the single "best" model, with dozens of publications to date providing examples.

Outside of the fishery science literature, advocacy of model averaging as a means of providing improved point estimates has an extensive history, starting as far back as a theoretical result by Laplace (1818, summarized by Stigler 1973) and a simple empirical demonstration by Galton (1907), and with increasing frequency in the 1960s-1980s literature on "combining forecasts" (reviewed by Clemen 1989). With the development of Bayesian model averaging (e.g., Chatfield 1995, Draper 1995), model averaging also garnered interest as a means of providing improved estimates of statistical uncertainty, and the rapid growth of literature on "machine learning" in the last few decades has added greatly to the suite of tools available for model averaging (e.g., Knox 2018).

When attempting to apply some sort of model averaging, two questions jump immediately to the forefront, and there is currently no scientific consensus on the answer to either: 1) How should the set of models be chosen, and 2) how should the average be weighted (recognizing that equal weighting is itself a form of weighting)? The importance of the first question notwithstanding, this paper will concern itself only with the second.

Within the stock assessment literature, many authors have suggested that model weights should ideally consist of Bayesian posterior probabilities. However, computation of such probabilities can be difficult and, perhaps more importantly, requires that the same data be used to fit all of the models in the set (e.g., Hill et al. 2007, Ianelli et al. 2016). Although some model averaging studies have successfully produced fully Bayesian probabilities for the models in the set (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), retrospective analysis (e.g., Rossi et al. 2019), or "true skill statistic" (e.g., Tanaka et al. 2020). Equal weighting of models has also been used (e.g., Stewart and Martell 2015, Ianelli et al. 2016, Rossi et al. 2019, Bryndum-Buchholz et al. 2020, Holsman et al. 2020, Reum et al. 2020).

Many model averaging approaches and applications arose in disciplines such as weather and climate forecasting, in which 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 ensemble. However, in stock assessment, a time series of true values for the quantity of interest is invariably lacking. For example, the assessment scientist seldom, if ever, has at his or her disposal a time series of true stock sizes or true harvest levels corresponding to a maximum sustainable yield (*msy*) exploitation rate. One possibility is to optimize the set of model 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.

In other words, for the stock assessment scientist, not only is the true model structure unknown, but the task is confounded further by the fact that there is no unambiguously true standard to which an ensemble can be tuned. This paper proposes a new method, cross-conditional model averaging (CCMA), which attempts to address these twin difficulties.

Methods

The basic construct in CCMA is an "ensemble of ensembles;" more specifically, a probability-weighted average of optimally weighted model averages (ensembles).

As noted above, two main types of model weights have been addressed in the literature, broadly speaking. The "combining forecasts" literature focuses on weights that optimize the performance of a linear combination of the models in a set; while Bayesian model averaging, and the various approximations thereto, focus on the probability that the structure of any given model within the set corresponds to the true structure, or (perhaps equivalently) on how well any given model, taken by itself, fits the data. The CCMA approach combines the two by using the former to specify the weights in each individual ensemble, and the latter to form a weighted average of the individual ensembles.

A more mathematically explicit description of the method follows. Because modern statistical methods of producing point estimates can typically be cast in terms of the distribution of the estimator (e.g., as the mean or median), this description will focus on deriving the distribution. The choice of point estimator will be addressed in the Discussion.

The general case

In principle, CCMA provides a distribution for some quantity of interest *z* as a weighted average of ensembles, where each ensemble is, in turn, a weighted average of distributions arising from *nmod* individual models, as follows:

$$f_{ave}(z) = \sum_{i=1}^{nmod} \left(p_i \sum_{j=1}^{nmod} W_{i,j}\, f\left(z \middle| m_i \otimes m_j\right) \right),$$

where **p** is a an *nmod* vector of probabilities (summing to unity); **W** is an *nmod*×*nmod* matrix of non-negative model weights (each row summing to unity), **m** is an *nmod* vector of model identifiers; *f* is the distribution of *z*; and the notation $m_i \otimes m_j$ means that the structure of model *i* is *conditionally* true, but the distribution is *nevertheless based* on the structure of model *j*. In the special case where **W** is the identity matrix, CCMA reduces to conventional model averaging.

In practice, of course, the individual *f* terms are unknown and must be approximated by use of data, so CCMA proceeds as follows:

Given an observed data vector **x** of length *nobs*, each model *i* generates an estimate of *z* as follows:

$$z_i = g(m_i \oplus \mathbf{x}),$$

where the notation $m_i \oplus \mathbf{x}$ means that model *i* is fit to **x**, and *g* maps the results from fitting model *i* to **x** into the quantity of interest *z*.

There are many ways of approximating the distribution of *z* from $m_i \oplus \mathbf{x}$ (e.g., inverting the Hessian matrix, use of Markov chain Monte Carlo methods (MCMC)). However, when cross-conditioning is involved, conditional parametric bootstrapping (Gavaris 1993, Smith et al. 1993) appears to be the most straightforward approach. After model *i* is fit to the observed data vector **x**, conditional parametric bootstrapping is used to generate *nrep* replicated data vectors of length *nobs*, resulting in an *nobs*×*nrep* data matrix **X**, written as $\mathbf{X}(m_i \oplus \mathbf{x})$ to emphasize its dependence on m_i and **x**. Then, a cross-conditionally estimated vector of *z* values is generated for each combination of *i*=1,2,...,*nmod* and *j*=1,2,...,*nmod* by evaluating the following for *k*=1,2,...,*nrep*:

$$\left(Z_{i,j}\right)_k = g\left(m_j \oplus \mathbf{X}(m_i \oplus \mathbf{x})^{\langle k \rangle}\right),$$

where **Z** is an *nmod*×*nmod* matrix, each element of which is an *nobs* vector of estimated *z* values; and the superscript $\langle k \rangle$ denotes the *k*th column of the matrix **X**.

Let $\boldsymbol{\zeta}$ represent the midpoints of *nbin* equally sized histogram bins spanning the range [min(**Z**),max(**Z**)] and let $h(\mathbf{z},\boldsymbol{\zeta})$ return the frequencies of a vector **z** within the set of equally sized histogram bins centered at $\boldsymbol{\zeta}$. Then, f_{ave} is approximated as follows:

$$f_{ave}(\boldsymbol{\zeta}) = \sum_{i=1}^{nmod} \left(p_i \sum_{j=1}^{nmod} W_{i,j}\, h\left(Z_{i,j}, \boldsymbol{\zeta}\right)/nrep \right).$$

As in Bayesian model averaging, each element of **p** should ideally represent the probability that the structure of the associated model is true, conditional on the assumption that the set of models includes the model with the true structure. However, as noted in the Introduction, computation of Bayesian posterior model probabilities may be prohibitive, either because of computational difficulties or, more problematically, because the models do not all use the same data. Therefore, it is anticipated that **p** will more commonly be set by one of the approximations listed in the Introduction.

In contrast, **W** is unique to cross-conditional model averaging, and has a precise statistical objective, viz., optimizing each cross-conditional weighted average ensemble. Specifically, for any given model *i* (the "pivot" model), the elements of the corresponding row of **W** are chosen, subject to the constraints of non-negativity and summation to unity, so as to minimize the mean squared error for the ensemble, conditional on the structure of the pivot model being true. Quantification of "error," of course, requires specification of an optimal estimate whenever the true value of *z* is unknown, as will typically be the case for most quantities of interest in a stock assessment. The optimal estimate of *z* for any given model *i* ($zopt_i$) is assumed to be either:

- z_i, the estimate obtained by fitting model *i* to the observed data **x**; or
- mean($Z_{i,i}$), the mean of the estimates of *z* obtained by fitting model *i* to the conditional parametric bootstrap data that were generated by fitting that model to the observed data **x**.

Given $zopt_i$, the mean (across data replicates) squared error for pivot model *i* is given by

$$mse_i\left(\mathbf{W}^{T\langle i\rangle}\right) = \frac{1}{nrep} \sum_{k=1}^{nrep} \left(\sum_{j=1}^{nmod} W_{i,j}\left(Z_{i,j}\right)_k - zopt_i \right)^2,$$

where the superscript *T* denotes matrix transpose.

Once each row of **W** has been optimized (subject to the constraints of non-negativity and summing to unity) by minimizing the above equation, thus yielding the best-performing ensemble conditional on the structure of the associated pivot model being true, the mean squared errors can be averaged (across pivot models) to give the *expected* mean squared error:

$$emse(\mathbf{W}) = \sum_{i=1}^{nmod} p_i mse_i\left(\mathbf{W}^{T\langle i\rangle}\right).$$

The *emse* can be used to compare the performance of the general CCMA approach to various special cases; for example, conventional model averaging (where **W** is the identity matrix), or any single-model approach (where each element of a given column of **W** is equal to 1 and all other elements are equal to 0).

The two-model special case

To facilitate exploration of CCMA, this paper makes use of a pair of simple, two-model examples.

In the special case where *nmod*=2, the fact that the elements of **p** and each row of **W** must sum to unity means that **p** can be expressed entirely in terms of a single element (i.e., $\mathbf{p} = [p_1 \quad 1-p_1]^T$), and **W** can be expressed entirely in terms of the diagonal vector **w** as follows:

$$\mathbf{W} = \begin{bmatrix} w_1 & 1-w_1 \\ 1-w_2 & w_2 \end{bmatrix}$$

For pivot model i =1,2, define the mean squared error as:

$$mse_i(w_i) = \frac{1}{nuse_i} \sum_{k=1}^{nuse_i} \left(\left(w_i \left(Z_{i,i}\right)_k + (1-w_i)\left(Z_{i,3-i}\right)_k \right) - zopt_i \right)^2$$

For vectors **x** and **y** of length n, let *mean*(**x**), *var*(**x**), and *cov*(**x**,**y**) represent the sample arithmetic mean, sample variance, and sample covariance, respectively (where all three use n as the denominator). Mean squared error can then be rewritten as a quadratic function of w_i by defining a 5×2 matrix of coefficients **C** from the moments of the elements of **Z** for i=1,2 as follows:

$$C_{1,i} = var\left(Z_{i,i} - Z_{i,3-i}\right) + mean\left(Z_{i,i} - Z_{i,3-i}\right)^2$$

$$C_{2,i} = var\left(Z_{i,3-i}\right) - mean\left(Z_{i,3-i}\right) mean\left(Z_{i,i} - Z_{i,3-i}\right) - cov\left(Z_{i,i}, Z_{i,3-i}\right)$$

$$C_{3,i} = mean\left(Z_{i,i} - Z_{i,3-i}\right)$$

$$C_{4,i} = mean\left(Z_{i,3-i}\right)$$

$$C_{5,i} = var\left(Z_{i,3-i}\right)$$

Given the above, the mean squared error can be rewritten for each pivot model i=1,2 as:

$$mse_i(w_i) = C_{1,i} {w_i}^2 - 2\left(C_{2,i} + C_{3,i} zopt_i\right) w_i + \left(C_{4,i} - zopt_i\right)^2 + C_{5,i}$$

The value of w_i that minimizes the mean squared error, subject to the constraint $0 \le w_i \le 1$, is then obtained straightforwardly as:

$$w_i = min\left(1, max\left(0, \frac{C_{2,i} + C_{3,i} zopt_i}{C_{1,i}}\right)\right)$$

Finally, the expected mean squared error (i.e., the weighted average of mse_i across pivot models i=1 and i=2) is given by

$$emse(\mathbf{w}) = p_1 mse_1(w_1) + (1-p_1) mse_2(w_2)$$

Some special cases of *emse* are as follow:

Approach	*emse*
Single model, using model 1	$emse([1 \quad 0]^T)$
Single model, using model 2	$emse([0 \quad 1]^T)$
Conventional model averaging	$emse([1 \quad 1]^T)$
Cross-conditional model averaging	$emse([w_1 \quad w_2]^T)$

Each of the examples described below involves two models. For both examples:

- The models are numbered 1 and 2, but a four-character label is also assigned to each model for ease of reference.
- Model 1 was chosen as the true model, from which a single "observed" data set was simulated.
- Parameters were estimated by the method of maximum likelihood.
- Simply for convenience, **p** was specified on the basis of AIC weights.
- Except where indicated otherwise, each $zopt_i$ was set equal to the base estimate.

Example 1: Order means as estimated by two distributional forms

For positive random variable x, probability density function $f(x)$, and real number q, the mean of order q is defined as the qth root of the qth noncentral moment (e.g., Mitrinović 1970):

$$m(q) = \left(\int_0^\infty f(x) x^q \, dx \right)^{1/q},$$

which, in the limit as $q \to 0$, converges to

$$m(0) = exp\left(\int_0^\infty f(x) ln(x) \, dx \right).$$

Well known examples of order means include the arithmetic ($q = 1$), geometric (obtained in the limit as $q \to 0$), and harmonic ($q = -1$) means.

For this example, the models were distinguished by the functional form of $f(x)$. Model 1 (label: *logn*) used the lognormal distribution (Aitchison and Brown 1957) and Model 2 (label: *invg*) used the inverse Gaussian distribution (Tweedie 1957).

Additional background information for this example is given in Appendix 1.

To conduct the simulations, the following dimensions were specified:

- *nobs* = 20
- *nrep* = 100,000

The true mean and true coefficient of variation were set at values of 10.0 and 2.0, respectively. True values and base estimates of the parameters are shown below ("True" is shown in quotes for the inverse Gaussian distribution because it is not the true distributional form, but the "true" values listed do correspond to the true mean and true coefficient of variation):

Lognormal			Inverse Gaussian		
Parameter	True	Estimate	Parameter	"True"	Estimate
μ	1.498	1.550	α	0.250	0.396
σ	1.269	1.134	β	10.000	9.444

Based on AIC weighting, p_1 was assigned a value of 0.448 in this example.

The CCMA approach was applied multiple times in this example, each time using an order mean ranging from $q = -5$ to $q = 5$ as the quantity of interest, for the purpose of understanding how the performance of the approach varies with the choice of quantity of interest.

Estimates of bias for the two possible choices of $zopt_i$ (z_i, the estimate obtained by fitting model i to the observed data **x**; or mean($Z_{i,i}$), the mean of the estimates of z obtained by fitting model i to the conditional parametric bootstrap data that were generated by fitting that model to the observed data **x**) were obtained by repeating the simulation 1000 times (i.e., generating a new observed data vector **x**) for each example value of q and measuring bias as mean($zopt_i$)/$m(q)-1$, where mean($zopt_i$) was computed across the 1000 repeated simulations.

Example 2: Stock assessment models with two forms for the stock-recruitment relationship

For the second example, two very simple stock assessment models were developed, distinguished only by the form of the stock-recruitment relationship. Model 1 (label: *beho*) used the form suggested by Beverton and Holt (1957), and Model 2 (label: *rick*) used the form suggested by Ricker (1954). The parameters of both models consist of:

- *umsy*, the discrete annual exploitation rate corresponding to maximum sustainable yield
- *v*, the discrete annual natural mortality rate
- x_0, the initial stock size
- *xmsy*, the stock size corresponding to maximum sustainable yield
- σ, the standard deviation of log-scale observation (measurement) error

In addition to the parameters, results will also be presented for two derived quantities:

- *xpro*, the projected stock size for the coming year
- *ofl*, the "overfishing level," defined as the product of *umsy* and *xpro*

Additional background information for this example is given in Appendix 2.

To conduct the simulations, the following dimensions were specified:

- *nobs* = 30
- *nrep* = 10,000

True values and base estimates of the parameters and derived quantities are shown below:

Parameter	true	*beho*	*rick*
umsy	0.200	0.227	0.237
v	0.200	0.192	0.033
x_0	30.769	29.856	30.187
xmsy	100.000	74.619	68.889
σ	0.200	0.191	0.188
xpro	81.712	71.350	69.926
ofl	16.342	16.176	16.569

Based on AIC weighting, p_1 was assigned a value of 0.417 in this example.

Results

Example 1: Order means as estimated by two distributional forms

Base estimates (i.e., the maximum likelihood estimates based on the observed data) of integer order means from –5 to 5 are shown for each of the two models below:

order:	–5	–4	–3	–2	–1	0	1	2	3	4	5
logn:	0.19	0.36	0.69	1.30	2.48	4.71	8.96	17.05	32.43	61.69	117.35
invg:	0.88	1.06	1.34	1.80	2.68	4.71	9.44	17.73	28.57	40.93	54.26

The inverse Gaussian model generally gave higher estimates than the lognormal for integer order means –5 through 2 (although the estimates for order 0 were essentially equal), while the lognormal model gave higher estimates than the inverse Gaussian for integer orders 3 through 5.

Figure 1 addresses the extent to which optimized model weights (**w**) were specific to the quantity being estimated, and the extent to which they mattered, by profiling over a range of order means from –5 to 5 (not restricted to integer values). As demonstrated by Figure 1, the optimal value of **w** can vary dramatically, depending on the quantity being estimated. For example, w_1=1 for all order means less than about –0.9, while w_1=0 for all order means greater than about 2.3; whereas w_2=1 for all order means between about –0.2 and 0.6 and greater than about 1.8, and w_2 never falls below about 0.4.

Figure 2 shows the proportional reductions in *emse* achieved by CCMA relative to the other three approaches (single-model using *logn*, single-model using *invg*, and conventional model averaging (*conv*)).

The improvement resulting from use of CCMA were sometimes fairly modest:

- For order means between about –1.2 and 1.8, improvements resulting from use of CCMA never exceeded about 10% relative to any of the other three alternatives.
- For order means greater than about 2.3, relative to the single-model approach using *invg*, use of CCMA did not result in any improvement.
- For order means between about –4.4 and 2.2, improvement relative to *conv* was less than 10%.

However, in other circumstances, the improvement was substantial. For example, reductions in excess of 50% were obtained by CCMA for:

- order means less than about –4.0 or greater than about 2.6, relative to the single-model approach using *logn*.

- order means less than about –3.0, relative to the single-model approach using *invg*.
- order means greater than about 3.0, relative to *conv*.

To examine a particular case in greater detail, Figure 3 shows results for $q = -5$, with *nbin*=50. The optimal values of w_1 and w_2 were 1.000 and 0.775, respectively. Curves are color-coded as follows: blue corresponds to Model 1 (*logn*), red corresponds to Model 2 (*invg*), and purple corresponds to some sort of model average (*conv* or *ccma*).

The upper two panels of Figure 3 correspond to each cross-conditional pair of model fittings to *nrep* sets of bootstrap data replicates from a conditionally true model. The upper left panel corresponds to the case where Model 1 (*logn*) is the conditionally true model, and the upper right panel corresponds to the case where Model 2 (*invg*) is the conditionally true model. In each of the upper panels, the dashed curve represents the conditionally true model, the dotted curve represents the other model, and the solid curve represents the optimized *ccma* ensemble for that cross-conditional pair (i.e., the average of the dashed and dotted curves, weighted by w_1 and $1-w_1$ in the case of the upper left panel, and by w_2 and $1-w_2$ in the case of the upper right panel). In the event that the respective element of **w** is 0.0 or 1.0, the solid curve will overlay the dotted or dashed curve, respectively.

The lower two panels of Figure 3 develop the final model averages. The lower left panel develops the *conv* average, and the lower right panel develops the *ccma* average. In the lower left panel, the dashed blue and dashed red curves duplicate their counterparts from the upper left (blue) and upper right (red) panels, and the dashed purple curve represents the *conv* average (i.e., the average of the dashed blue and red curves, weighted by p_1 and $1-p_1$). In the lower right panel, the solid blue and solid red curves duplicate their counterparts from the upper left (blue) and upper right (red) panels, and the solid purple curve represents the *ccma* average (i.e., the average of the solid blue and red curves, weighted by p_1 and $1-p_1$).

Means and standard deviations of the distributions shown in Figure 3 are provided below (where the notation $i \otimes j$ means that the bootstrap replicates were generated by model i and fit by model j), and *ccma*(i) represents the cross-conditional ensemble for the case where model i is conditionally true):

Statistic	1⊗1	1⊗2	2⊗1	2⊗2	*ccma*(1)	*ccma*(2)	*ccma*	*conv*
mean:	0.344	0.987	0.300	0.999	0.344	0.841	0.618	0.705
sdev:	0.326	0.411	0.260	0.320	0.326	0.424	0.456	0.459

As expected, both the cross-conditional and conventional model averaging approaches (*ccma* and *conv*) reflect greater uncertainty in the estimated quantity than either single model by itself (1⊗1 and 2⊗2).

Relative changes in *emse* for the $q = -5$ case were as follow (row relative to column):

m(–5)	*logn*	*invg*	*conv*	*ccma*
logn	0	-0.3391	1.2875	1.5963
invg	0.5131	0	2.4611	2.9284
conv	-0.5628	-0.7111	0	0.1350
ccma	-0.6148	-0.7454	-0.1189	0

Of the two single-model approaches, *logn* outperformed *invg* (i.e., it gave a lower *emse*). Both model-averaging approaches outperformed both of the single-model approaches, and, of the two model-averaging approaches, *ccma* outperformed *conv* (by about 12%).

Finally, Figure 4 addresses the issue of which choice of $zopt_i$ (the base estimate z_i or mean($Z_{i,i}$)) is superior, by showing the bias in base estimates and sample means from each model for integer order means ranging from –5 to 5. Both estimators (base and mean) for both models were close to unbiased for order means from –1 to 1. However, performance of either estimator, for both models, degrades considerably for order means outside this range. For the *logn* model, the mean estimator performed much more poorly than the base estimator outside the [–1,1] range, whereas for the *invg* model, the performance of the mean estimator was only slightly worse than that of the base estimator throughout the range.

Example 2: Stock assessment models with two forms for the stock-recruitment relationship

Of the model fits to the *nrep*=10,000 bootstrap data replicates generated for each pivot model, some were determined to be unreliable, either due to a large gradient (>0.0001) or a Hessian matrix that did not appear to be positive definite. In the event that a model failed to provide a satisfactory fit to a bootstrap data replicate, the results from both models (for that data replicate) were discarded. As a result, results from 994 of the bootstrap data replicates were discarded when Model 1 was the pivot model and results from 803 of the bootstrap data replicates were discarded when Model 2 was the pivot model. Thus, wherever *nrep* appears in the description of the CCMA approach in the Methods section, it was replaced by $nuse_1$=9,006 and $nuse_2$=9,197 here.

It should be noted that the results for *xmsy* contained several extremely large estimates. In the interest of completeness, results for *xmsy* will be reported here, even though some of those results may not be particularly useful (e.g., the mean and standard deviation of the distribution seem unreasonably large).

The optimized values of **w** for the parameters and derived quantities were as follow (values of 0 and 1 are highlighted red and green, respectively, for easy identification):

Quantity	w_1	w_2
umsy	0.7227	0.3002
v	0.7810	0.1161
x_0	1.0000	0.0000
xmsy	0.5319	0.0909
σ	1.0000	0.0000
xpro	0.1177	1.0000
ofl	0.7292	0.2863

As with Example 1, the results for Example 2 demonstrate that the optimal value of **w** can vary dramatically depending on the quantity being estimated.

Table 1 shows the relative changes in *emse* for each quantity and all combinations of estimator. In nearly all cases, the *emse* for *ccma* was less than that for any other estimator. The two exceptions were the two cases where one element of **w** was 0.0 and the other was 1.0, in which case *ccma* was tied with one of the single-model approaches (specifically, the single-model approach using *beho* in both cases). Relative to *conv*, the reductions in *emse* achieved by *ccma* were sometimes small (e.g., 1% to 2% for *xpro* and σ), but were also sometimes quite substantial (e.g., >40% for *v*, *umsy*, *ofl*, and *xmsy*; although the *xmsy* result may be suspect, as noted above). The *conv* approach always outperformed one of the single-model approaches but not the other: *conv* gave a lower *emse* than the single-model approach using *beho* in the case of *xpro*, and a lower *emse* than the single-model approach using *rick* in all other cases. Of the six cases where the single-model approach using *beho* outperformed the *conv* approach, some of the improvements were substantial (e.g., the reduction in *emse* for the single-model approach using *beho*,

relative to the *emse* for *conv*, exceeded 30% for *umsy*, *v*, *xmsy*, and *ofl*; although, again, the *xmsy* result may be suspect).

Probability mass functions, following the format and conventions of Figure 3, are shown for the parameters and derived quantities in Figure 5, where *nbin*=50 in each case (for the panels associated with *xmsy*, each horizontal axis has been truncated so that it corresponds to the maximum of the upper end of the 67% concentrations, due to the extremely heavy tails of the full distributions).

Means and standard deviations of the distributions shown in Figure 5 are provided in Table 2.

Discussion

As shown here, when performance is measured in terms of expected mean squared error, CCMA can perform no worse than either the best single-model approach or conventional model averaging, and the results from the two simple examples demonstrate that CCMA can sometimes perform substantially better than either of these (in contrast, note that conventional model averaging can actually perform worse than a single-model approach).

Although the method has yet to be applied to a set of stock assessment models involving more than a few parameters each, the theory described here and the results from the two simple examples suggest that CCMA is worthy of further exploration as a tool for improving stock assessment estimates. In anticipation of such exploration, some possible methodological issues, potential subjects for future research, and practical considerations for implementation are addressed below.

Possible methodological issues

Use of bootstrap distributions as an approximation of Bayesian posterior distributions

In CCMA, at least as developed to date, 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 could be found for obtaining the Bayesian posterior distributions required by CCMA, it would probably be preferable to use such an approach (subject to computational feasibility), but this appears difficult, because of the need to develop distributions for each model that are conditional on the structure of a given pivot model being true.

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, the 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 distributions 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 on 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 CCMA addresses such potential over-weighting explicitly, by requiring each model to take a turn as the operating (pivot) model.

Given the facts that: 1) all fishery modeling involves approximations, 2) 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 CCMA is likely not a critical flaw.

Assuming that the model set contains the true model

The assumption that the model set contains the true model, while almost surely invalid in the strict sense, is widely made in the model selection literature, particularly the 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)). Moreover, even when the impossibility of ever identifying the true model is acknowledged, the fact remains that, in practice, the use of stock assessment results in a fishery management context typically involves proceeding *as though* one of the models in the model set is the true model (e.g., Kass and Raftery 1995). In CCMA, this assumption is required in order to take the final step of converting the individual optimized ensembles, each of which is conditional on the structure of the respective pivot model being true, into an (unconditional) expectation across all pivot models. Thus, the assumption that the model set contains the true model is best viewed as a matter of convenience (or logical necessity) rather than a matter of ontology.

Potential subjects for future research

Relation of CCMA to "combining forecasts"

As noted in the Methods section, the CCMA approach bears some similarity to techniques described in the "combining forecasts" literature (Clemen 1989), to which Reid (1968) and Bates and Granger (1969) provided foundational contributions. Granger and Ramanathan (1984) noted that the standard techniques for combining forecasts tended to take the form of linear regression models, with the observed value as the dependent variable and the individual model forecasts as the independent variables. They noted that, while existing applications typically fixed the intercept term at zero and constrained the model weights so that they sum to unity (as is done for each row of **W** in CCMA), such constraints are not logically necessary, and that relaxing those constraints might be expected to result in better performance of the combined forecast. However, Phillips (1987) showed that the constrained solution can sometimes actually outperform the unconstrained solution, and further noted that, even when the unconstrained solution performs better, the performance of the constrained solution can be very nearly as good. Whether removal of the constraints on **W** could be expected to improve the performance of CCMA is suggested as a subject for future research.

Relation of CCMA to "machine learning"

Machine learning techniques have been used in several fishery applications of ensemble modeling (e.g., Walsh and Kleiber 2001, Soykan et al. 2014, Liu et al. 2020, and Siders et al. 2020). To the non-practitioner, the terminology of machine learning can be confusing, with some techniques being so closely related that it is difficult to distinguish between them meaningfully. One machine learning

technique that relates to CCMA is "bagging" (shorthand for "bootstrap aggregation," Breiman 1996a). Both bagging and CCMA employ bootstraps, but bagging typically uses *un*conditional *non*parametric bootstraps (i.e., the original bootstrapping method of Efron 1979; see also the classification of bootstrapping methods proposed by Smith et al. 1993), whereas CCMA uses conditional parametric bootstraps.

The CCMA approach also appears to be closely related to "stacking" (Wolpert 1992, Breiman 1996b) and the "super learner" (van der Laan et al. 2007) and "superensemble" (Krishnamurti et al. 1999; see Anderson et al. 2017 and Rosenberg et al. 2018 for stock assessment applications) techniques. In all of those techniques, the models are fit to the observed data and used to estimate the quantity of interest, and then those estimates are treated as variables to be used in fitting the ensemble to the observed data by tuning the model weights, akin to the "combining forecasts" technique discussed above. A major difference between these machine learning techniques and CCMA is that the former use observed data only, so no cross-conditioning is possible and the "ensemble of ensembles" aspect of CCMA is absent (note that some authors, e.g., Sipper and Moore (2021), use the term "superensemble" to refer to an ensemble of ensembles, but this is different from the superensemble technique of Krishnamurti et al.). Another difference is that these machine learning techniques all involve cross-validation in an attempt to avoid over-fitting. Note, however, that cross-validation is not typically used in bagging, because the use of bootstrapping is thought to serve the same purpose. Further clarification of the relation between CCMA and existing machine learning techniques, and exploration of whether the performance of CCMA would be improved by incorporating a cross-validation step even though it already employs bootstrapping, are also suggested as subjects for future research.

Point estimation

As noted in the Introduction, both point estimates and estimates of uncertainty are important, and model averaging has been suggested as a means of improving both. The description of CCMA in the Methods section was focused on deriving the distribution of the quantity of interest, anticipating that the point estimate would be based on that distribution (e.g., as the mean or median).

Interpreted in the context of decision theory, the use of minimum squared error to estimate the elements of any given row of $\mathbf{W}$ implies the use a particular loss function. One of the characteristics of the loss function used here is that the optimal estimator is the arithmetic mean of the average distribution. With respect to the choice of measure used to represent $zopt_i$, and noting that the point estimate z_i will often be expected to approximate mean($Z_{i,i}$), the fact that the arithmetic mean of the average distribution is the optimal estimator suggests that the choice might not be terribly consequential. This conclusion is corroborated by the results from Example 1, in which the choice made little difference when estimating the population mean of order 1.

However, other choices for the loss function are possible. Appendix 3 describes a general loss function, of which the one used here is a special case. In the general form of the loss function, the optimal estimator is the mean of order $1-ra$, where ra is an *ad hoc* measure of risk aversion, so that the arithmetic mean of the estimated distribution is obtained as the optimal estimator (as here) only in the special case where $ra = 0$ (i.e., "risk neutrality"). There is a coherence between the order means of the distributions corresponding to the individual ensembles and the order mean of the average distribution; *viz*., the order mean of the average distribution equals the order mean of the individual order means (Appendix 3). This means that, for values of ra substantially different from 0, the sample mean of of order $1-ra$ (i.e., of the sample $Z_{i,i}$) might be a better choice to serve as $zopt_i$ than the point estimate z_i. This conclusion is also corroborated by the results from Example 1, in which both z_i and mean($Z_{i,i}$) were substantially biased when estimating means of order less than -1 or greater than 1.

Use of the general loss function in Appendix 3 is therefore suggested as another subject for future research.

Practical considerations for implementation

Choosing which models to include

As noted in the Introduction, one of the two fundamental problems involved in model averaging is choosing the set of models. Although solving this problem is outside the scope of the present paper, it should be acknowledged that the results from any type of model averaging can be influenced by the choices involved in establishing the set of models. Because each row of **W** is optimized for the ensemble specific to the pivot model associated with that row, CCMA presumably compensates to some extent for inclusion of duplicative, superfluous, or poorly performing models. However, because specification of **p** is not as straightforward (see below), there is still potential for results to be biased by a less-than-careful choice of models to include in the set. One strategy that may help to guard against subconscious "stacking of the deck" is to structure the set of models by some sort of factorial design.

Specification of model probabilities

The CCMA approach requires specification of the vector of model probabilities (**p**) used to compute the final expectation in the algorithm (i.e., the expectation across the set of pivot-specific, optimally weighted ensembles). Except (perhaps) for the fact that the vector of model probabilities carries a particular interpretation in the CCMA approach (*viz.*, **p** represents the probabilities of the various model structures in the set being the true model structure), the problem of specifying **p** in CCMA is similar to the problem of specifying model weights in conventional model averaging. As noted in the Introduction, there is currently no scientific consensus on how this should be accomplished in conventional model averaging, and the task is no easier in CCMA. Although an abundance of methods is available, choosing one will inevitably require making some subjective judgments.

Optimal intra-ensemble weighting depends of the quantity being estimated

The fact that the optimal estimate of **W** depends on the quantity being estimated, as shown here by both Examples 1 and 2, also raises a practical consideration. One possible course of action would be to optimize **W** for each quantity being estimated. In a stock assessment involving hundreds of parameters, this might appear daunting at best. However, in a superensemble described by Krishnamurti et al. (2016), as many as 10 million weights were estimated in order to optimize performance fully across all models, variables, time steps, and spatial grid-points. Alternatively, **W** could be optimized for the single quantity of greatest importance (e.g., *ofl*), then use that same **W** for estimation of all other quantities, recognizing that the resulting estimates for those other quantities would likely not be fully optimal (but at least they would be estimated consistently).

Some data subsets may not be used by all models

Another practical consideration arises if some data subsets are not used by all of the models. For example, it may be the case that the data set used by model M1 includes subset D1, while model M2 does not use subset D1 at all, meaning that, when model M2 takes its turn as the pivot model, it will not be possible to generate bootstrap values for subset D1 that are truly conditional on model M2 being the true model. A straightforward solution, however, would be to generate the bootstrap replicates for subset D1 based on that subset's sampling distribution. For example, survey index data might be developed from a sampling design that gives rise to estimates of normal or lognormal sample mean and variance parameters, and compositional data might be assumed to follow a multinomial distribution with sample

size equal to the number of hauls from which the samples were taken. In the worst-case scenario where the parameters of the sampling distribution cannot be estimated, it might be necessary to remove model M1 or M2 from the ensemble.

Time limitations

A significant practical consideration for implementing CCMA is the amount of time required to conduct the analysis. If the ensemble involves many models, or a few very complicated models, application of CCMA could potentially take a long time, as $nrep \times nmod^2$ runs are required. This could require adjusting expectations regarding use of the most recent data when conducting a stock assessment.

References

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

Aitchison, J., and J. A. C. Brown. 1957. *The Lognormal Distribution*. Cambridge University Press, Cambridge.

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. Rosenberg, and E. R. Selig. 2017. Improving estimates of population status and trend with superensemble models. *Fish Fish.* 18:732-741.

Bates, J. M., and C. W. J. Granger. 1969. The combination of forecasts. *Oper. Res. Q.* 20:451-468.

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

Beverton, R. J. H., and S. J. Holt. 1957. On the Dynamics of Exploited Fish Populations. *Fishery Investigations*, Series 2, Volume 19.

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.

Breiman, L. 1996a. Bagging predictors. *Machine Learning* 24:123-140.

Breiman, L. 1996b. Stacked regressions. *Machine Learning* 24:49-64.

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

Bryndum-Buchholz, A., D. G. Boyce, D. P. Tittensor, V. Christensen, D. Bianchi, and H. K. Lotze. 2020. Climate-change impacts and fisheries management challenges in the North Atlantic Ocean. *Mar. Ecol. Prog. Ser.* 648:1-17.

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.

Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. *J. R. Statist. Soc. A* 158:419-466.

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

Clemen, R. T. 1989. Combining forecasts: A review and annotated bibliography. *Int. J. Forecast.* 5:559-583

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.

Draper, D. 1995. Assessment and propagation of model uncertainty. *J. R. Statist. Soc. B* 57:45-97.

Efron, B. 1979. Bootstrap methods: another look at the jacknife. *Ann. Stat*. 7:1-26.

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.

Galton, F. 1907. Vox populi. *Nature* 75:450-451.

Gavaris, S. 1993. Analytical estimates of reliability for the projected yield from commercial fisheries. *In* S. J. Smith, J. J. Hunt, and D. Rivard (editors), Risk evaluation and biological reference points for fisheries management. *Can. Spec. Publ. Fish. Aquat. Sci*. 120:185-191.

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.

Granger, C. W. J., and R. Ramanathan. 1984. Improved methods of combining forecasts. *J. Forecast.* 3:197-204.

Hilborn, R. 2003. The state of the art in stock assessment: where we are and where we are going. *Sci. Mar.* 67(Suppl. 1):15-20.

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.

Holsman, K. K., A. C. Haynie, A. B. Hollowed, J. C. P. Reum, K. Aydin, A. J. Hermann, W. Cheng, A. Faig, J. N. Ianelli, K. A. Kearney, and A. E. Punt. 2020. Ecosystem-based fisheries management forestalls climate-driven collapse. *Nat. Commun*. 11:4579.

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.

Knox, S. W. 2018. *Machine Learning: A Concise Introduction*. John Wiley and Sons. Hoboken, NJ, USA. 331 p.

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.

Krishnamurti, T. N., V. Kumar, A. Simon, A. Bhardwaj, T. Ghosh, and R. Ross. 2016. A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Rev. Geophys.* 54:336-377.

Laplace, P. S. de. 1818. *Deuxième Supplément a la Théorie Analytique des Probabilités* Paris: Courcier. Reprinted (1847) in *Oeuvres de Laplace* 7:569-623. Paris: Imprimerie Royale; also reprinted (1886) in *Oeuvres Complètes de Laplace* 7:531-580. Paris: Gautheir-Villars.

Liu, C., S. Zhou, Y.-G. Wang, and Z. Hu. 2020. Natural mortality estimation using tree-based ensemble learning models. *ICES J. Mar. Sci*. 77:1414-1426.

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.

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

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. K. 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. 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.

Phillips, R. F. 1987. Composite forecasting: an integrated approach and optimality reconsidered. *J. Bus. Econ. Stat.* 5:389-395.

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.

Reid, D. J. 1968. Combining three estimates of gross domestic product. *Economica* 35:431-444.

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.

Reum, J. C. P., J. L. Blanchard, K. K. Holsman, K. Aydin, A. B. Hollowed, A. J. Hermann, W. Cheng, A. Faig, A. C. Haynie, and A. E. Punt. 2020. Ensemble projections of future climate change impacts on the eastern Bering Sea food web using a multispecies size spectrum model. *Front. Mar. Sci*. 7, Article 124.

Ricker, W.E. 1954. Stock and recruitment. *J. Fish. Res. Board Can.* 11:559-623.

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. 349–382.

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.

Siders, Z. A., N. D. Ducharme-Barth, F. Carvalho, D. Kobayashi, S. Martin, J. Raynor, T. T. Jones, and R. N. M. Ahrens. 2020. Ensemble random forests as a tool for modeling rare occurrences. *Endang. Species Res.* 43:183-197.

Sipper, M., and J. H. Moore. 2021. Conservation machine learning: a case study of random forests. *Sci. Rep.* 11:3629.

Smith, S. J., J. J. Hunt, and D. Rivard. 1993. Introduction. *In* Smith, S. J., J. J. Hunt, and D. Rivard. (editors), Risk evaluation and biological reference points for fisheries management. *Can. Spec. Publ. Fish. Aquat. Sci*. 120:vi-viii.

Soykan, C. U., T. Eguchi, S. Kohin, and H. Dewar. 2014. Prediction of fishing effort distributions using boosted regression trees. *Ecol. Appl*. 24:71-83.

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.

Stigler, S. M. 1973. Studies in the history of probability and statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* 60:439-445.

Tanaka, K. R., M. P. Torre, V. S. Saba, C. A. Stock, and Y. Chen. 2019. An ensemble of high-resolution projection of changes in the future habitat of American lobster and sea scallop in the Northeast US continental shelf. *Divers. Distrib*. 26:987-1001.

Tweedie, M. C. K. 1957. Statistical properties of inverse Gaussian distributions. *Ann. Math. Stat.* 28:362-377.

van der Laan, M. J., E. C. Polley, and A. E. Hubbard. 2007. Super learner. *Stat. Appl. Gen. Mol. Biol.* 6:1-23.

Walsh, W. A., and P. Kleiber. 2001. Generalized additive model and regression tree analyses of blue shark (Prionace glauca) catch rates by the Hawaii-based commercial longline fishery. *Fish. Res*. 53:115-131.

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.

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5:241-259.

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. Relative changes in expected mean squared error (row relative to column).

umsy	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.5168	-0.3520	0.2713
rick	1.0696	0	0.3411	1.6311
conv	0.5432	-0.2544	0	0.9618
ccma	-0.2134	-0.6199	-0.4903	0

v	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.5489	-0.4315	0.0553
rick	1.2167	0	0.2602	1.3393
conv	0.7590	-0.2065	0	0.8563
ccma	-0.0524	-0.5725	-0.4613	0

x_0	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.2220	-0.1332	0.0000
rick	0.2854	0	0.1142	0.2854
conv	0.1536	-0.1025	0	0.1536
ccma	0.0000	-0.2220	-0.1332	0

xmsy	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.5249	-0.5044	0.4500
rick	1.1050	0	0.0433	2.0522
conv	1.0176	-0.0415	0	1.9256
ccma	-0.3104	-0.6724	-0.6582	0

σ	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.0324	-0.0197	0.0000
rick	0.0335	0	0.0131	0.0335
conv	0.0201	-0.0129	0	0.0201
ccma	0.0000	-0.0324	-0.0197	0

xpro	*beho*	*rick*	*conv*	*ccma*
beho	0	0.0414	0.0296	0.0416
rick	-0.0398	0	-0.0113	0.0002
conv	-0.0287	0.0115	0	0.0117
ccma	-0.0400	-0.0002	-0.0115	0

ofl	*beho*	*rick*	*conv*	*ccma*
beho	0	-0.5210	-0.3605	0.2375
rick	1.0879	0	0.3353	1.5838
conv	0.5636	-0.2511	0	0.9350
ccma	-0.1919	-0.6130	-0.4832	0

Table 2. Means and standard deviations of distributions generated by conditional parametric bootstraps.

Quantity	Statistic	1⊗1	1⊗2	2⊗1	2⊗2	*ccma* (1)	*ccma* (2)	*ccma*	*conv*
v	mean	0.251	0.398	0.226	0.335	0.283	0.239	0.257	0.300
	sdev	0.297	0.398	0.261	0.380	0.328	0.279	0.301	0.350
umsy	mean	0.231	0.424	0.221	0.387	0.284	0.271	0.276	0.322
	sdev	0.248	0.300	0.217	0.272	0.277	0.247	0.260	0.274
x_0	mean	29.116	28.899	29.000	28.896	29.116	29.000	29.048	28.987
	sdev	2.424	2.768	2.387	2.690	2.424	2.387	2.403	2.585
xmsy	mean	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	sdev	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
σ	mean	0.178	0.177	0.177	0.176	0.178	0.177	0.177	0.177
	sdev	0.024	0.024	0.024	0.024	0.024	0.024	0.024	0.024
xpro	mean	72.537	72.116	72.120	71.526	72.166	71.526	71.793	71.948
	sdev	7.729	7.678	7.456	7.406	7.685	7.406	7.530	7.559
ofl	mean	16.836	30.630	16.017	27.805	20.572	19.391	19.883	23.231
	sdev	18.356	22.229	16.038	20.310	20.423	18.168	19.149	20.255

Figures

Figure 1. Model weights for a range of order means (Example 1).

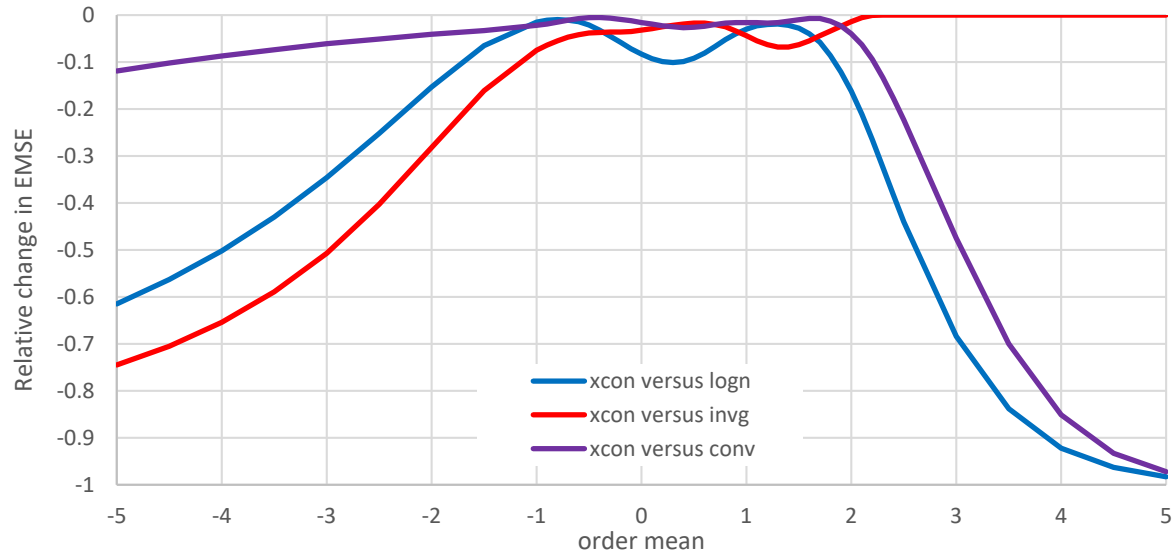

Figure 2. Expected mean squared errors for a range of order means (Example 1).

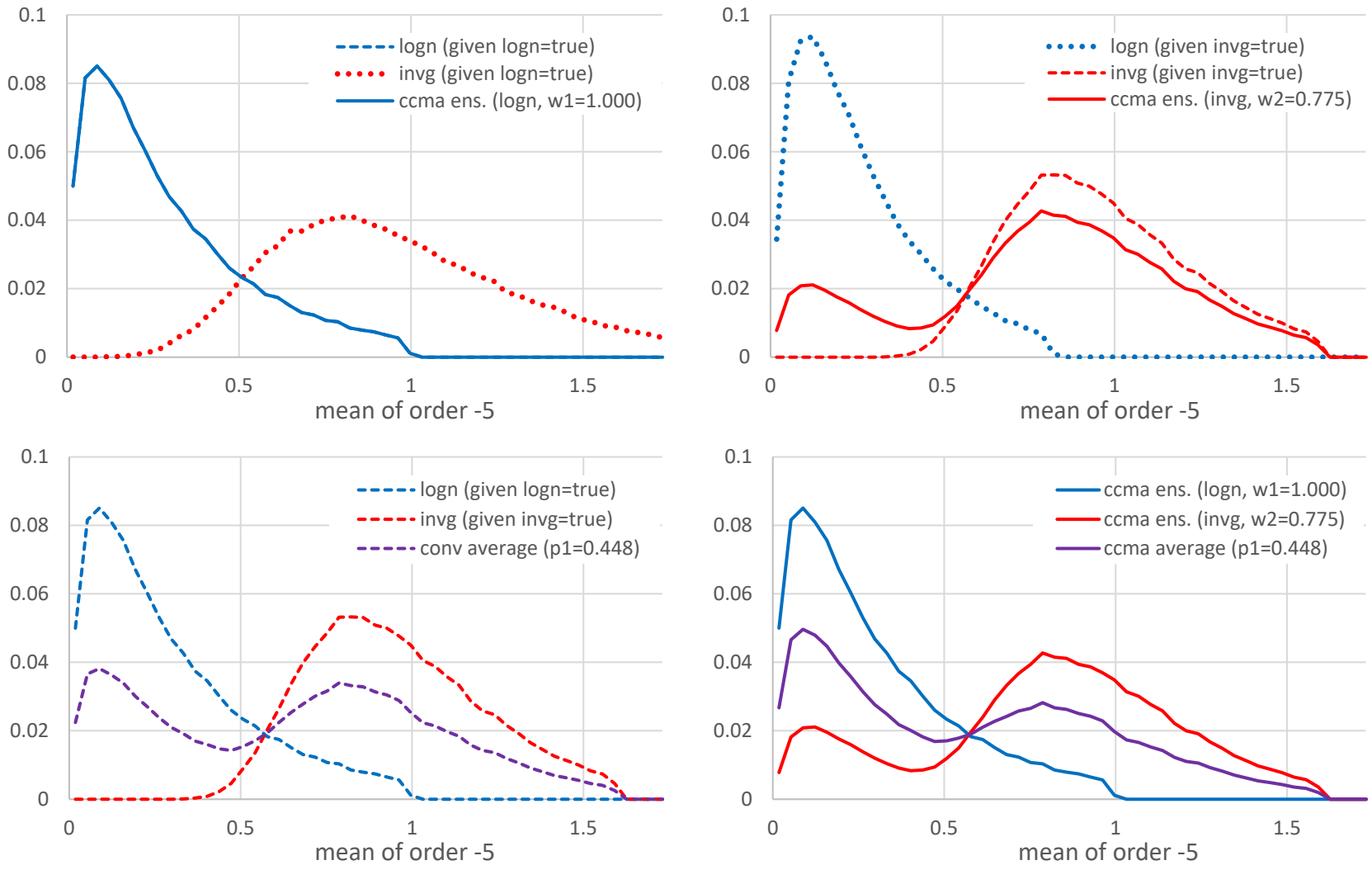

Figure 3. Probability mass functions generated by conditional parametric bootstraps (Example 1).

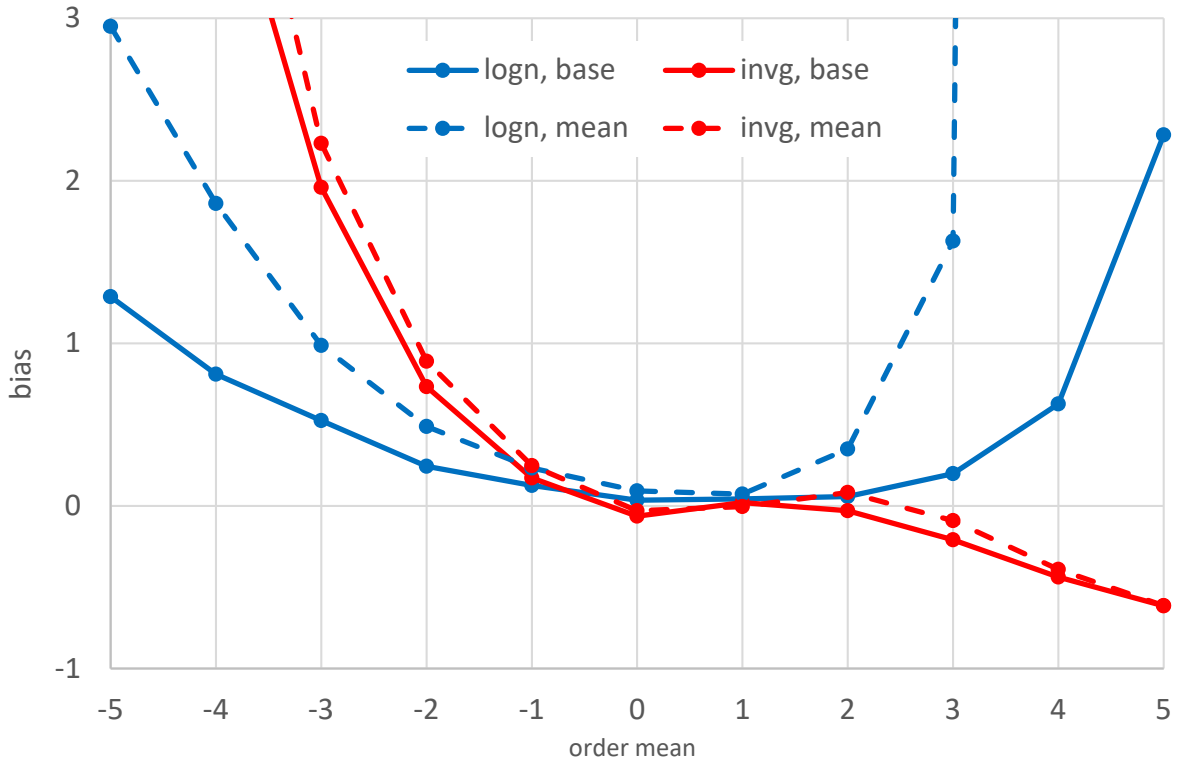

Figure 4. Bias in estimators for a range of order means (Example 1).

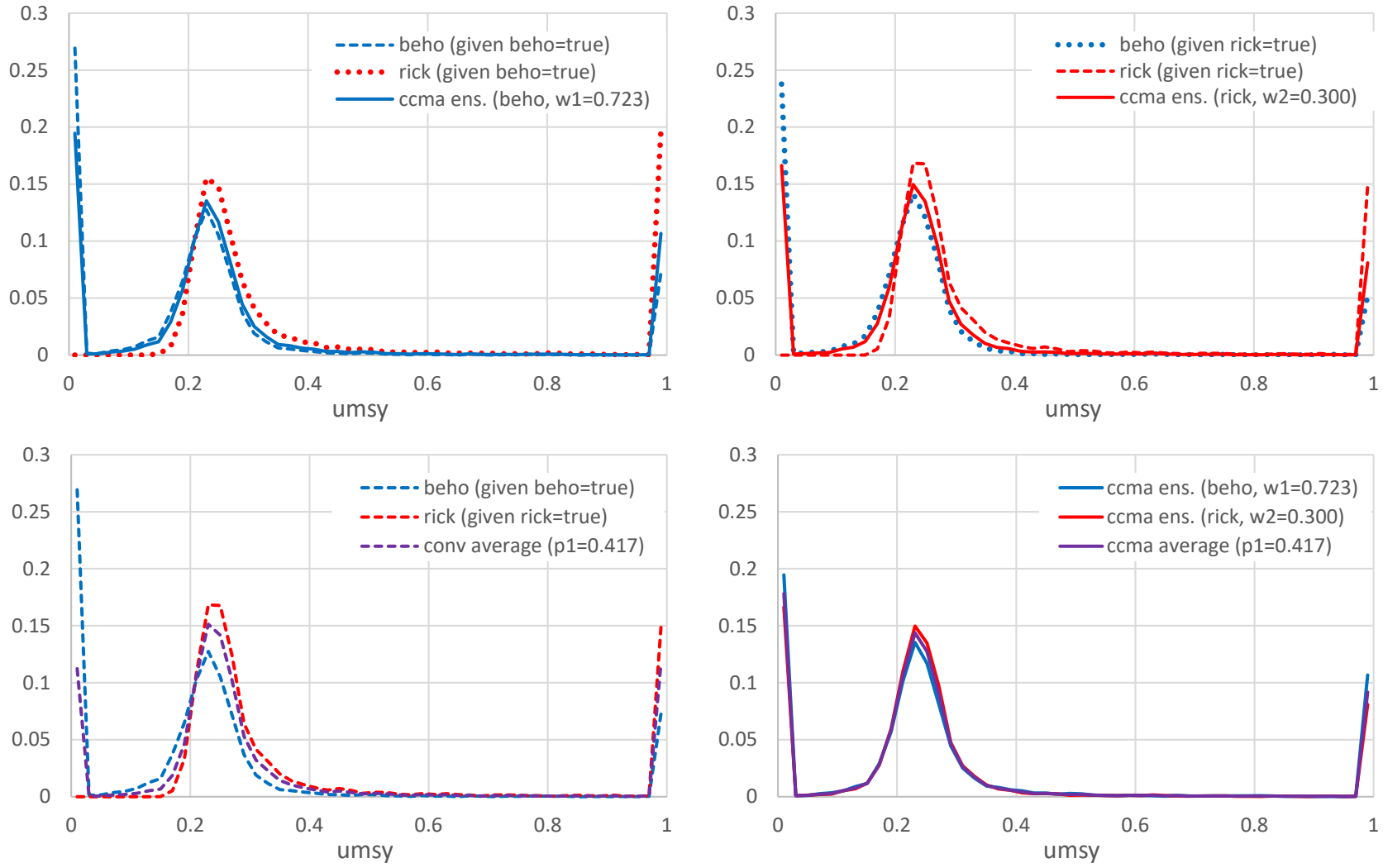

Figure 5a. Probability mass functions of *umsy* generated by conditional parametric bootstraps (Example 2).

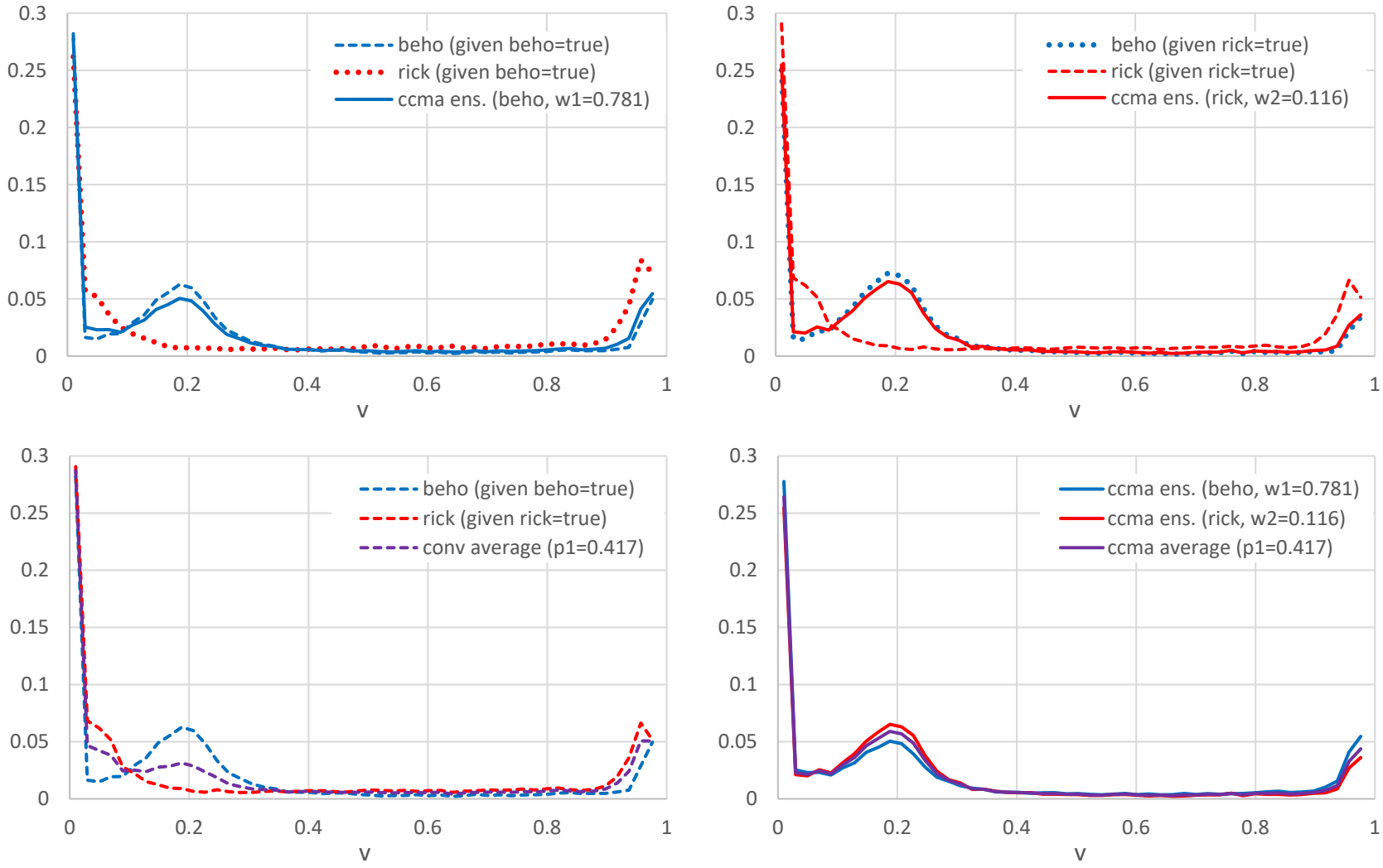

Figure 5b. Probability mass functions of *v* generated by conditional parametric bootstraps (Example 2).

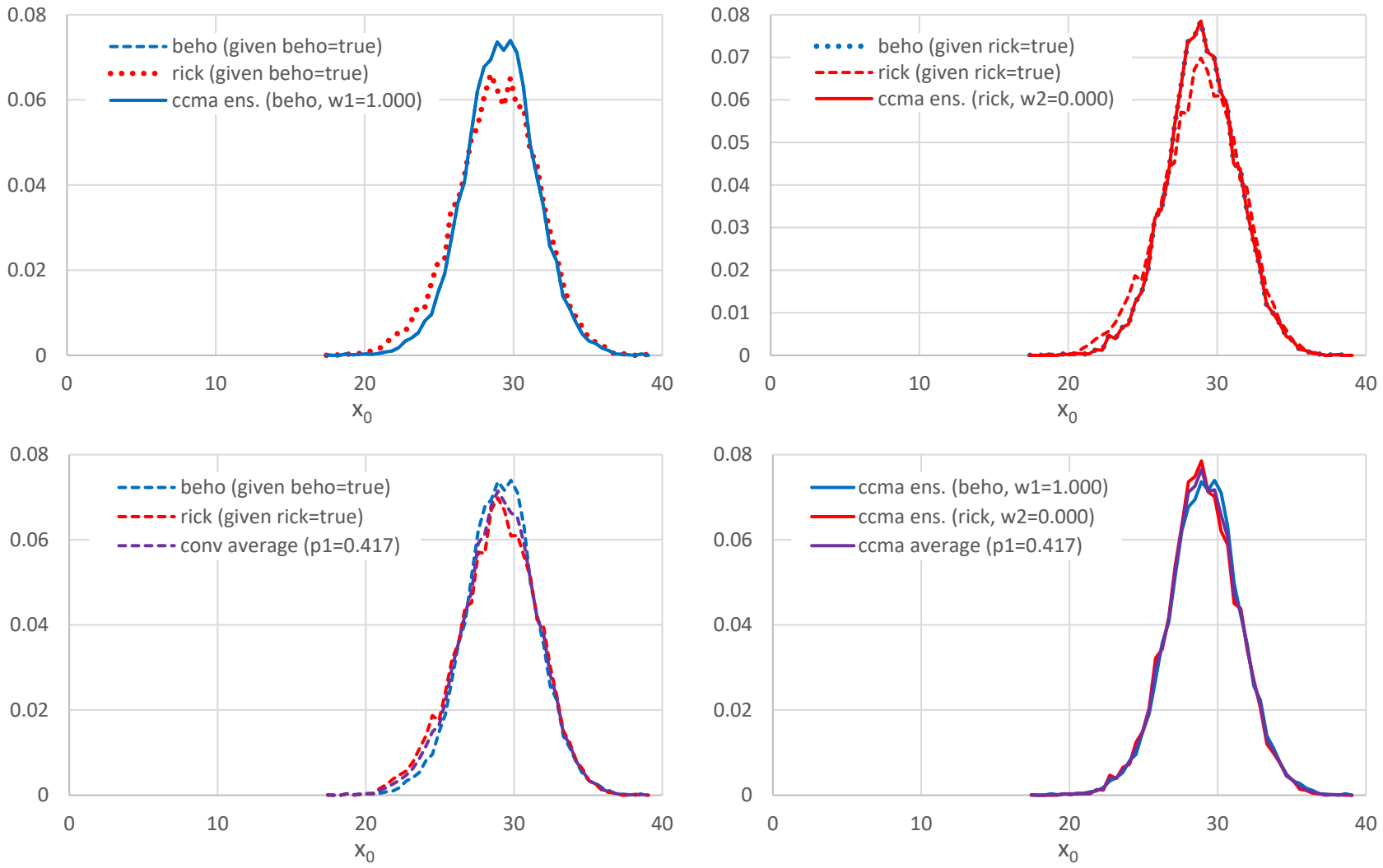

Figure 5c. Probability mass functions of x_0 generated by conditional parametric bootstraps (Example 2).

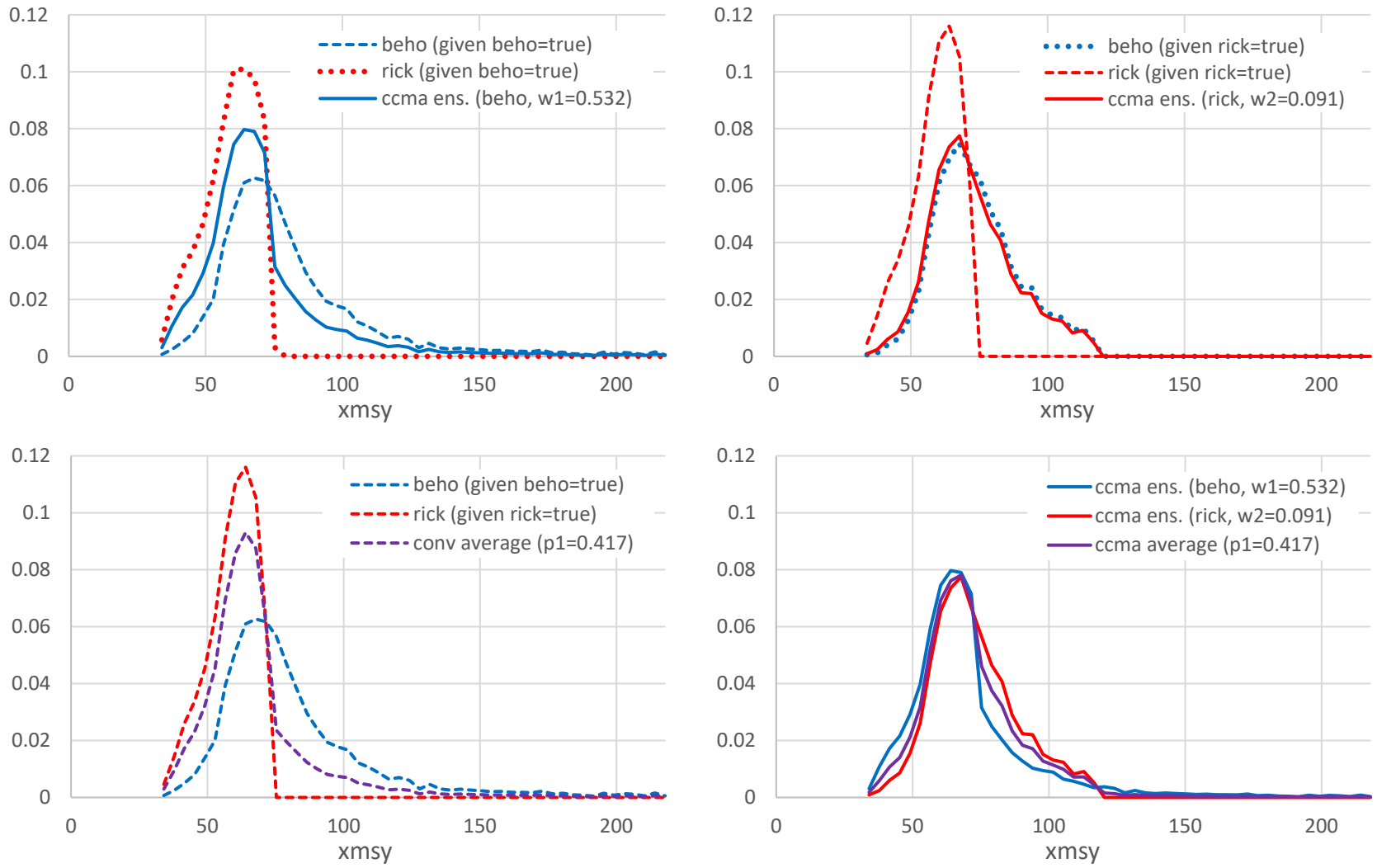

Figure 5d. Probability mass functions of *xmsy* generated by conditional parametric bootstraps (Example 2).

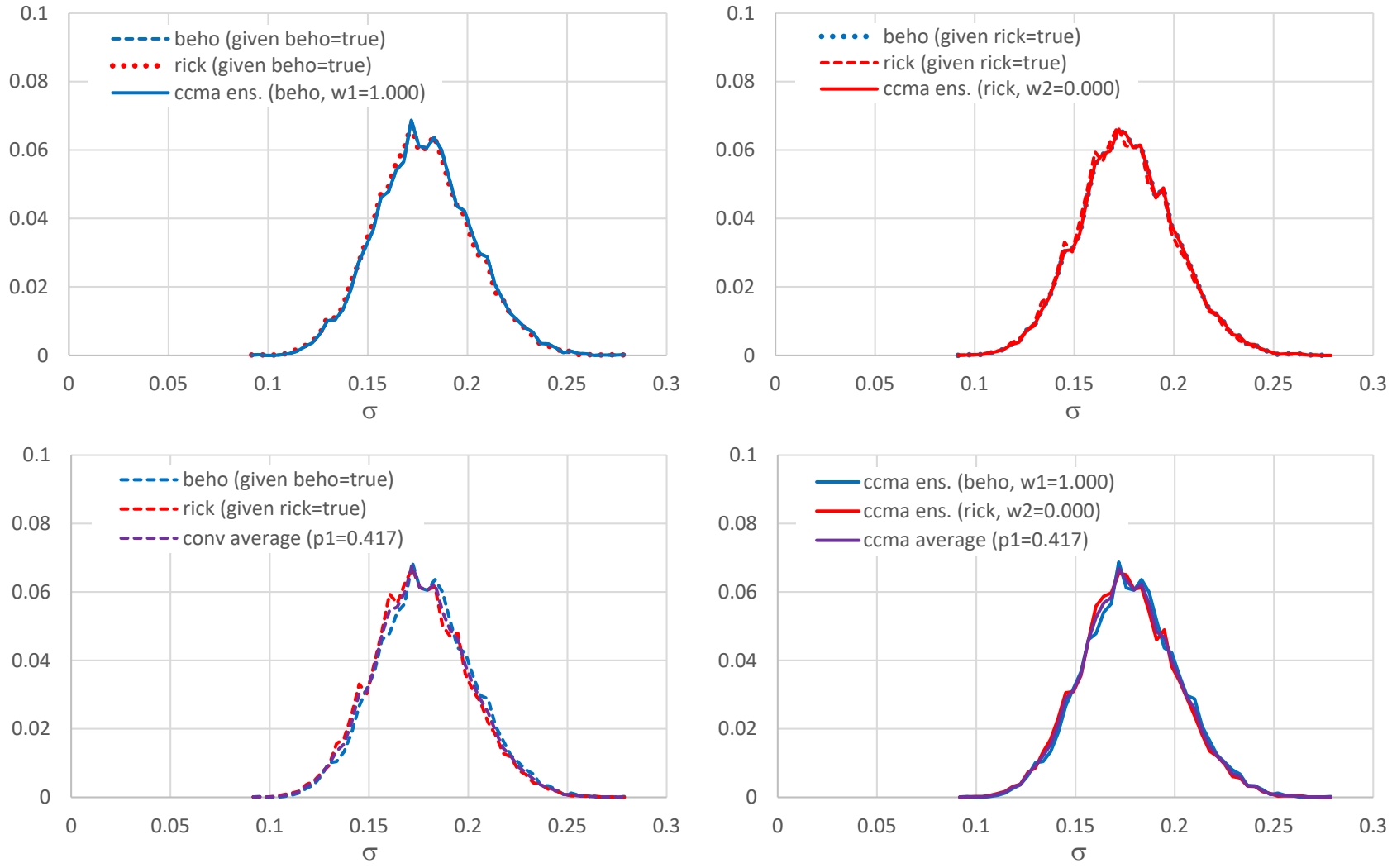

Figure 5e. Probability mass functions of σ generated by conditional parametric bootstraps (Example 2).

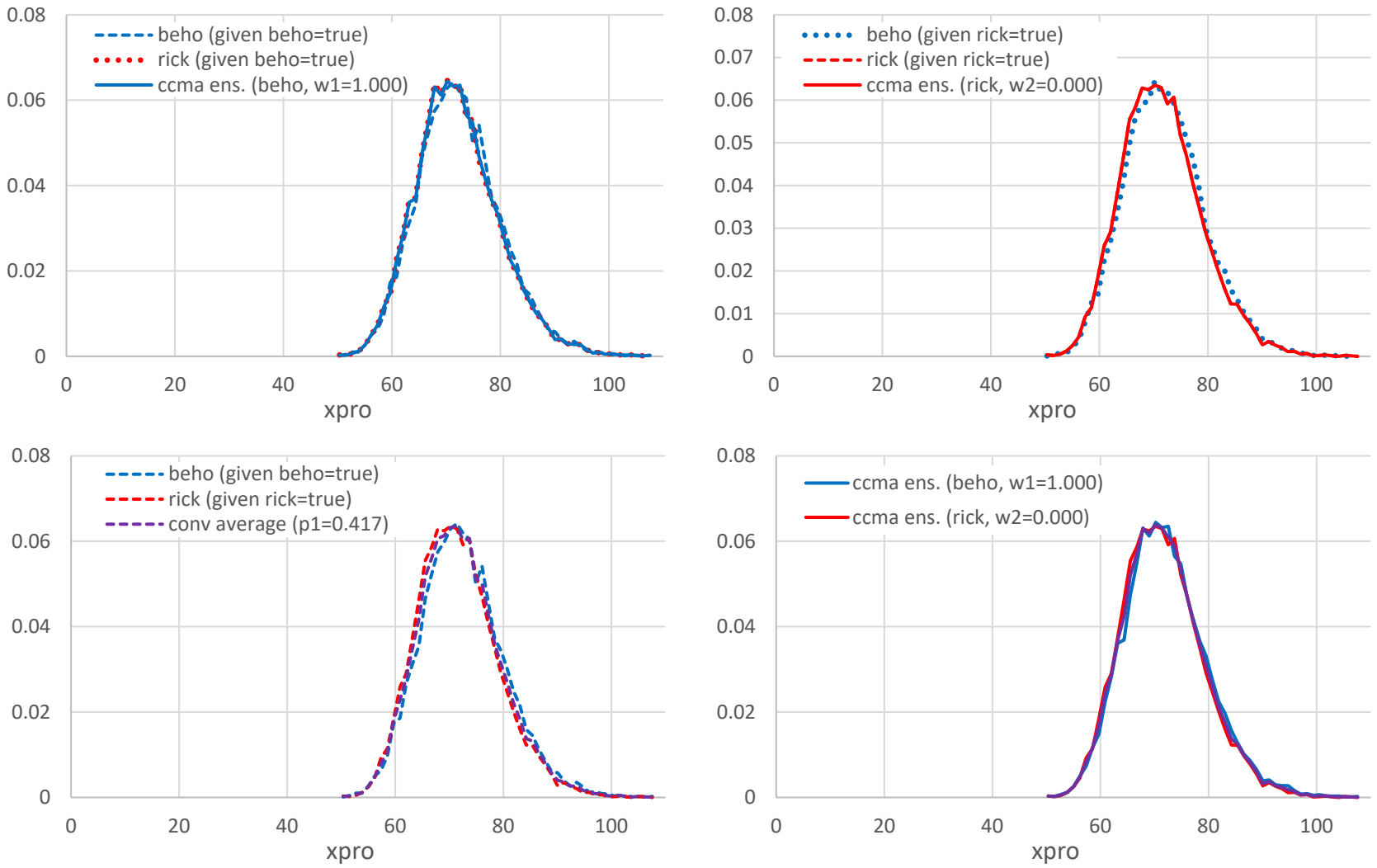

Figure 5f. Probability mass functions of *xpro* generated by conditional parametric bootstraps (Example 2).

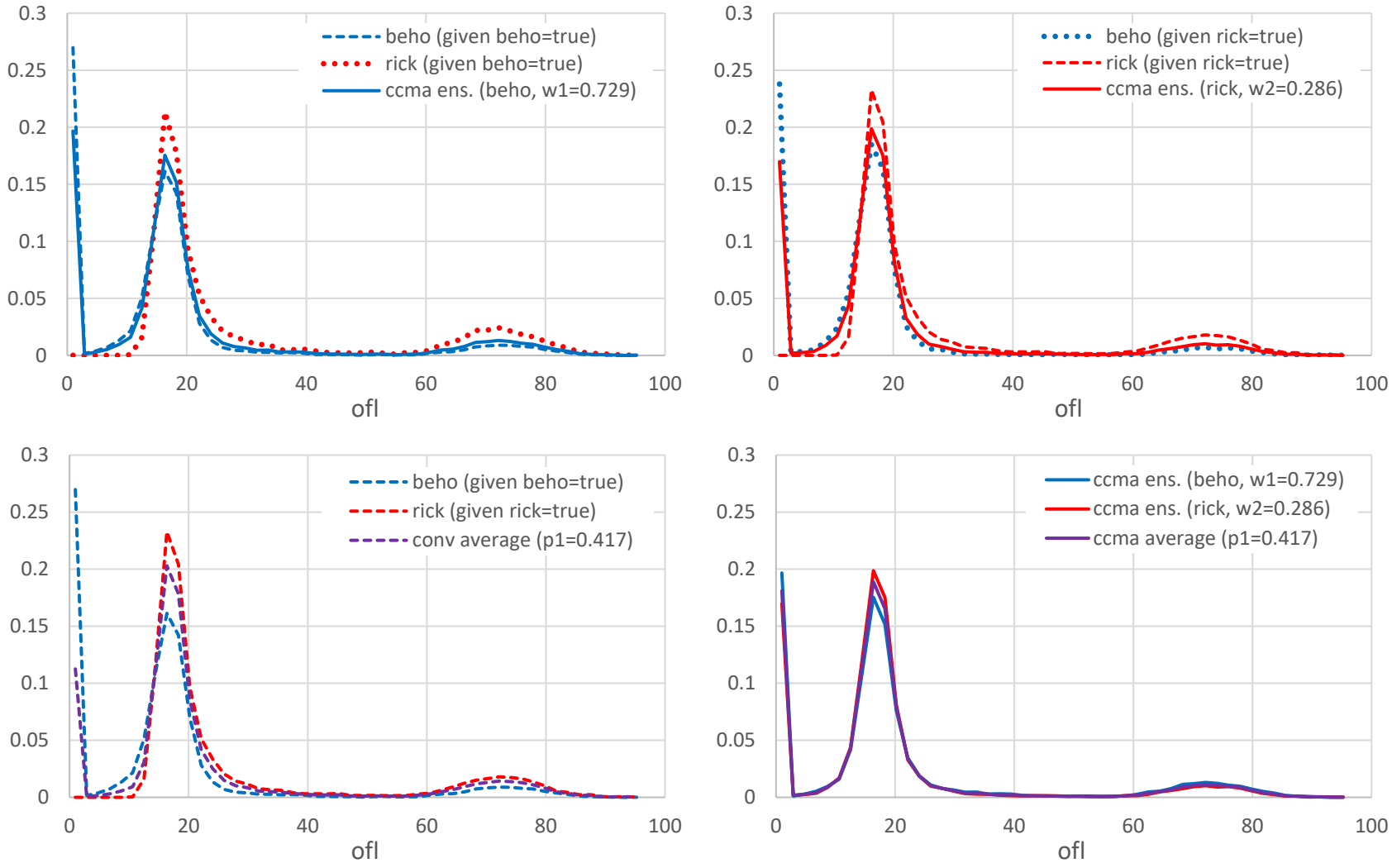

Figure 5g. Probability mass functions of *ofl* generated by conditional parametric bootstraps (Example 2).

Appendix 1

Additional background information for the distributional example

The lognormal distribution (label: *logn*) can be written as

$$logn(x) = \sqrt{\frac{1}{2\pi}}\left(\frac{1}{\sigma x}\right)\exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right),$$

and the inverse Gaussian distribution (label: *invg*) can be written as

$$invg(x) = \frac{\left(\frac{x}{\beta}\right)^{-3/2}\exp\left(-\frac{\alpha}{2}\left(\frac{x}{\beta}+\frac{\beta}{x}\right)\right)}{\beta exp(-\alpha)\sqrt{2\pi/\alpha}}.$$

If the two distributions share the same mean and coefficient of variation (*cv*), their shapes are nearly indistinguishable by eye so long as *cv* is sufficiently small (e.g., less than about 0.5; Figure A1.1). For larger values of *cv*, the inherently greater skewness of the lognormal distribution (by a factor of $1 + cv^2/3$) becomes apparent. The two distributions also share the characteristic that, for a given *cv*:

$$\frac{m(-1)}{m(1)} = \frac{1}{1+cv^2}.$$

The order means for the lognormal distribution are given by

$$m_{logn}(q) = \exp\left(\mu + \frac{q\sigma^2}{2}\right),$$

and for the inverse Gaussian distribution by

$$m_{invg}(q) = \beta\left(\exp(\alpha)\sqrt{\frac{2\alpha}{\pi}}K\left(\left|q-\frac{1}{2}\right|,\alpha\right)\right)^{1/q},$$

where $K(n,\cdot)$ is the *n*th-order modified Bessel function of the second kind (Abramowitz and Stegun 1972). Note that the order means for integer values of *q* (except zero) can be written without resorting to use of Bessel functions:

$$m(q) = \beta\left(\sum_{j=0}^{-q}\frac{(-q+j)!}{j!\,(-q-j)!\,(2\alpha)^j}\right)^{1/q} \quad \text{for } q < 0,$$

$$m(q) = \beta\left(\sum_{j=0}^{q-1}\frac{(q-1+j)!}{j!\,(q-1-j)!\,(2\alpha)^j}\right)^{1/q} \quad \text{for } q > 0.$$

The maximum likelihood estimates (MLEs) of the parameters for the two distributions are given by:

$$\begin{bmatrix}\hat{\mu}\\ \hat{\sigma}\end{bmatrix} = \begin{bmatrix}mean\left(ln(\mathbf{x})\right)\\ \sqrt{var\left(ln(\mathbf{x})\right)}\end{bmatrix} \text{ and } \begin{bmatrix}\hat{\alpha}\\ \hat{\beta}\end{bmatrix} = \begin{bmatrix}\dfrac{har(\mathbf{x})}{mean(\mathbf{x}) - har(\mathbf{x})}\\ mean(\mathbf{x})\end{bmatrix},$$

where *har*(**x**) is the harmonic mean of **x**.

Below are the cases where the MLE of the *population* order mean equals the *sample* order mean:

order	*logn*	*invg*
1	no	yes
0	yes	no
–1	no	yes

However, it should be note that, in general, the sample geometric and harmonic means are positively biased estimators of their respective true population counterparts.

Specifications for the experiment:

- *nobs* = 20
- *nrep* = 100,000
- true mean = 10
- true cv = 2
- true distributional form: lognormal

The randomly drawn data for the experiment are shown below:

2.562	1.889	2.453	1.337	0.527	4.726	3.838	9.059	72.128	12.477
15.606	13.353	14.288	10.503	1.189	4.882	1.715	10.823	3.551	1.975

Reference

Abramowitz, M., and I. A. Stegun. 1972. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (9th printing). Dover, New York.

Figure

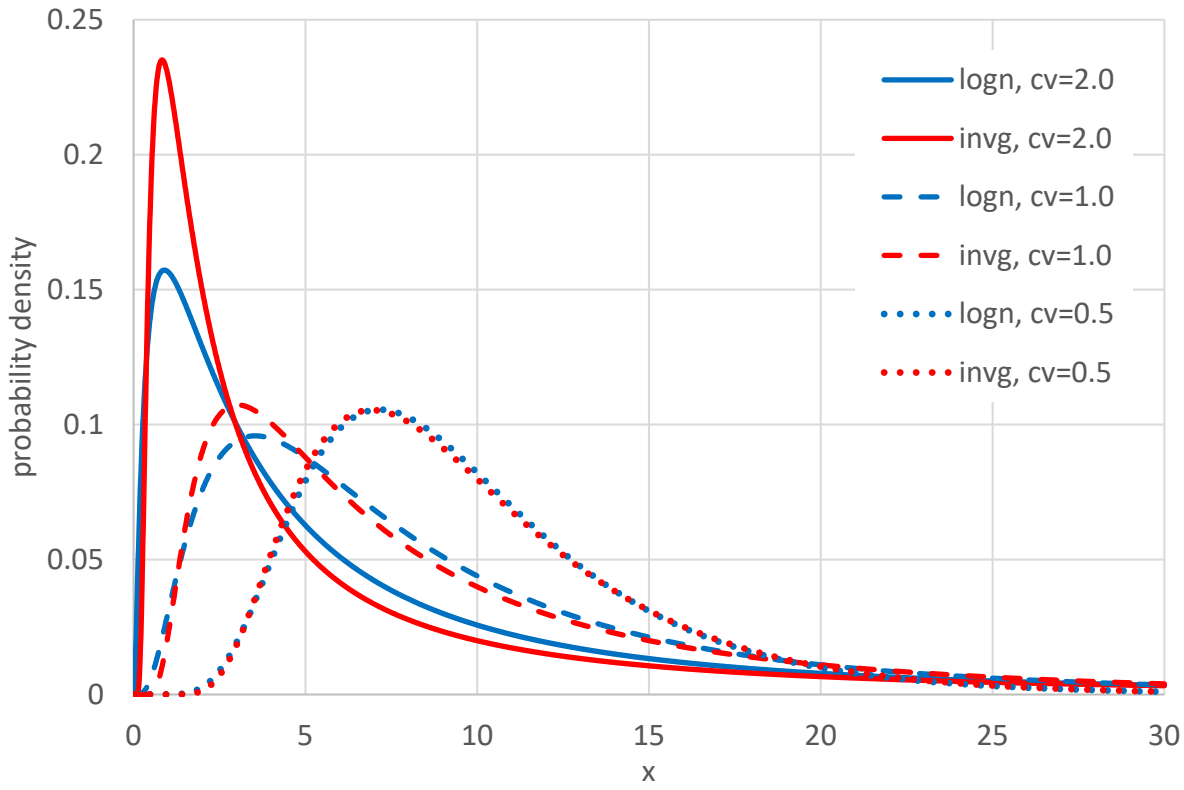

Figure A1.1. Example lognormal and inverse Gaussian pdfs.

Appendix 2

Additional background information for the stock assessment example

Two simple stock assessment models were structured with the following features in common:

- Stock dynamics are entirely deterministic
- Time *t* is measured in units of integer years, ranging from 0 to *ntim*
- Recruitment *r* in year *t* is determined by stock size in year *t*–1
- Discrete annual exploitation rate *u* varies linearly with *t* from u_0 to u_{ntim}
- Discrete annual natural mortality rate *v* is constant across *t*
- Stock size *x* is measured, with lognormal(0,σ^2) error, in units of fish density
- Initial stock size x_0 is in equilibrium at exploitation rate u_0
- Catch y_t is measured, without error, in units of fish density
- Intra-annual processes occur in the following order:
    - Measurement of stock size (except no measurement in year 0)
    - Recruitment
    - Exploitation
    - Natural mortality

True stock size and observed stock size for *t*=1,2,...,*ntim* can thus be described as follows:

$$x_t = x_{t-1} + r(x_{t-1}) - (u_{t-1}(1-v)+v)x_{t-1} \text{ and}$$

$$xobs_t = x_t \exp(\varepsilon_t),$$

where $\varepsilon_t \sim normal(0, \sigma^2)$, and catch for *t*=0,1,2,...,*ntim* is given by $y_t = u_t x_t$.

The only feature distinguishing the models was that Model 1 (label: *beho*) used a Beverton-Holt stock-recruitment relationship and Model 2 (label: *rick*) used a Ricker stock-recruitment relationship, each of which can be parameterized in terms of the slope at the origin (*a*) and a scale parameter (*b*) as follows:

$$r_{beho}(x) = \frac{a_{beho} b_{beho} x}{b_{beho} + x} \text{ and } r_{rick}(x) = a_{rick} x \exp\left(-\frac{x}{b_{rick}}\right).$$

Following a logic analogous to that of Schnute and Kronlund (1996), the parameters *a* and *b* for each model can be expressed in terms of *v* and the exploitation rate and stock size corresponding to maximum sustainable yield, *umsy* and *xmsy*, as:

$$\begin{bmatrix} a_{beho} \\ b_{beho} \end{bmatrix} = \begin{bmatrix} \frac{\lambda^2}{v} \\ xmsy\left(\frac{v}{\lambda - v}\right) \end{bmatrix} \text{ and}$$

$$\begin{bmatrix} a_{rick} \\ b_{rick} \end{bmatrix} = \begin{bmatrix} \exp\left(\frac{\lambda - v}{\lambda}\right)\lambda \\ xmsy\left(\frac{\lambda}{\lambda - v}\right) \end{bmatrix},$$

where $\lambda = umsy(1-v) + v$.

The following parameters were estimated statistically by maximum likelihood:

- *umsy* (logit transformed)
- *v* (logit transformed)
- x_0 (log transformed)
- *xmsy* (log transformed)
- σ (log transformed)

Note that it was unnecessary to estimate the annual exploitation rates u_t statistically, because the assumptions of deterministic dynamics and errorless measurement of catch meant that they were determined automatically.

To conduct the simulation, the following values were specified:

- *nrep* = 10,000
- *ntim* (= *nobs*) = 30
- *umsy* = 0.2
- u_0 = 0.4
- u_{ntim} = 0.2
- *v* = 0.2
- *xmsy* = 100
- σ = 0.2

The resulting stock-recruitment relationships for the two models, with total mortality lines corresponding to *u*=0 and *u*=*umsy*, are shown in Figure A2.1a, and the resulting sustainable yield curves for the two models are shown in Figure A2.1b.

Model 1 (*beho*) was assumed to be the true model, which, given the parameters specified above, resulted in the following set of derived quantities:

- initial stock size (x_0) = 30.769
- projected stock size in year *ntim*+1 (*xpro*) = 81.712
- overfishing level (*ofl*, the product of *umsy* and *xpro*) = 16.342

The following set of random normal error terms was generated:

time:	1	2	3	4	5	6	7	8	9	10
error:	0.066	-0.102	0.248	0.014	0.269	-0.131	-0.051	-0.249	0.109	-0.006
time:	11	12	13	14	15	16	17	18	19	20
error:	-0.019	0.046	0.034	-0.169	0.167	-0.016	-0.173	0.467	0.512	0.224
time:	21	22	23	24	25	26	27	28	29	30
error:	-0.179	0.007	0.362	-0.018	-0.018	-0.324	0.019	0.064	-0.159	-0.234

The time series of true and observed stock size are shown in Figure A2.2.

Reference

Schnute, J. T., and A. R. Kronlund. 1996. A management oriented approach to stock recruitment analysis. *Can. J. Fish. Aquat. Sci.* 53:1281-1293.

Figures

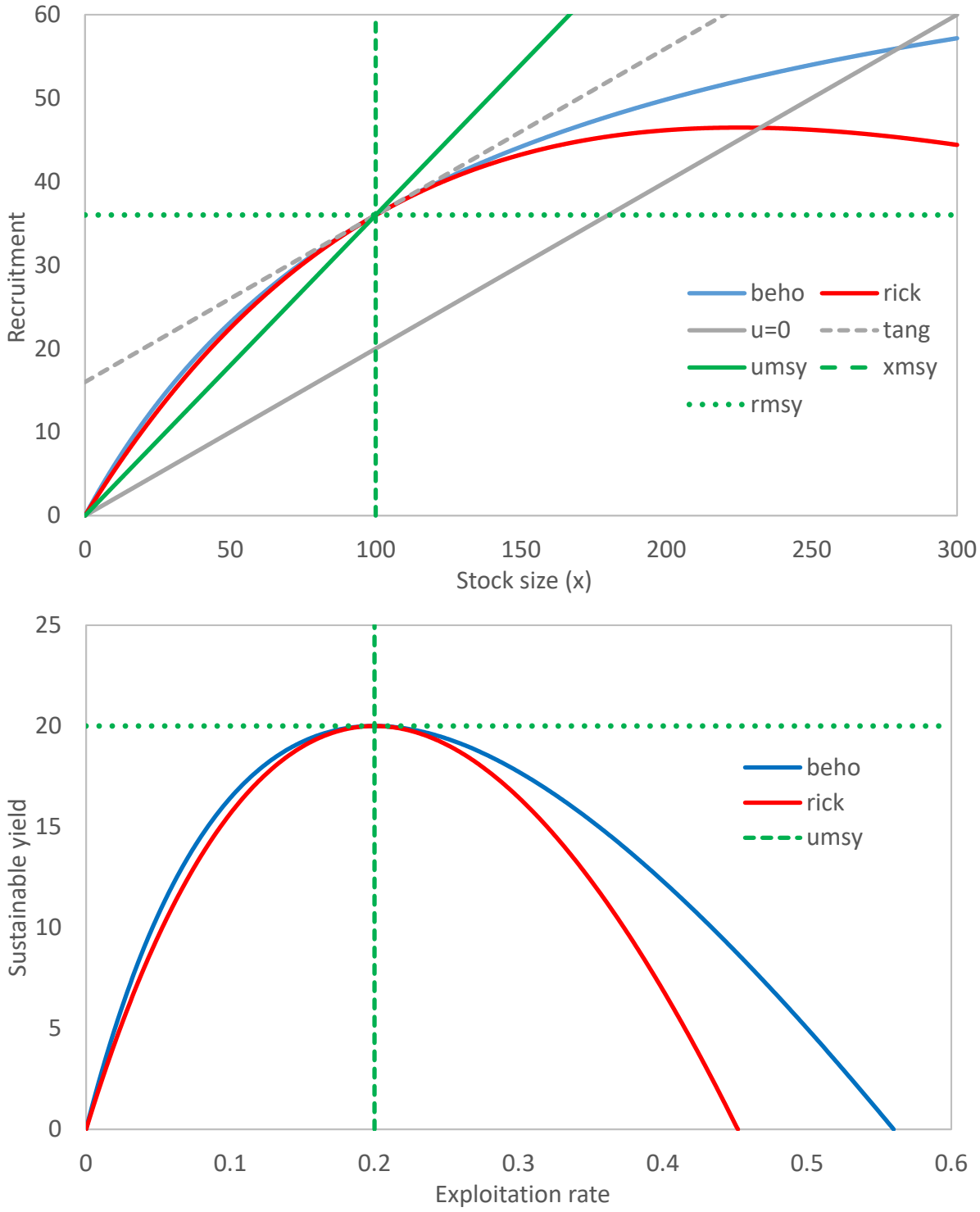

Figure A2.1. Equilibrium behavior of the *beho* and *rick* models.

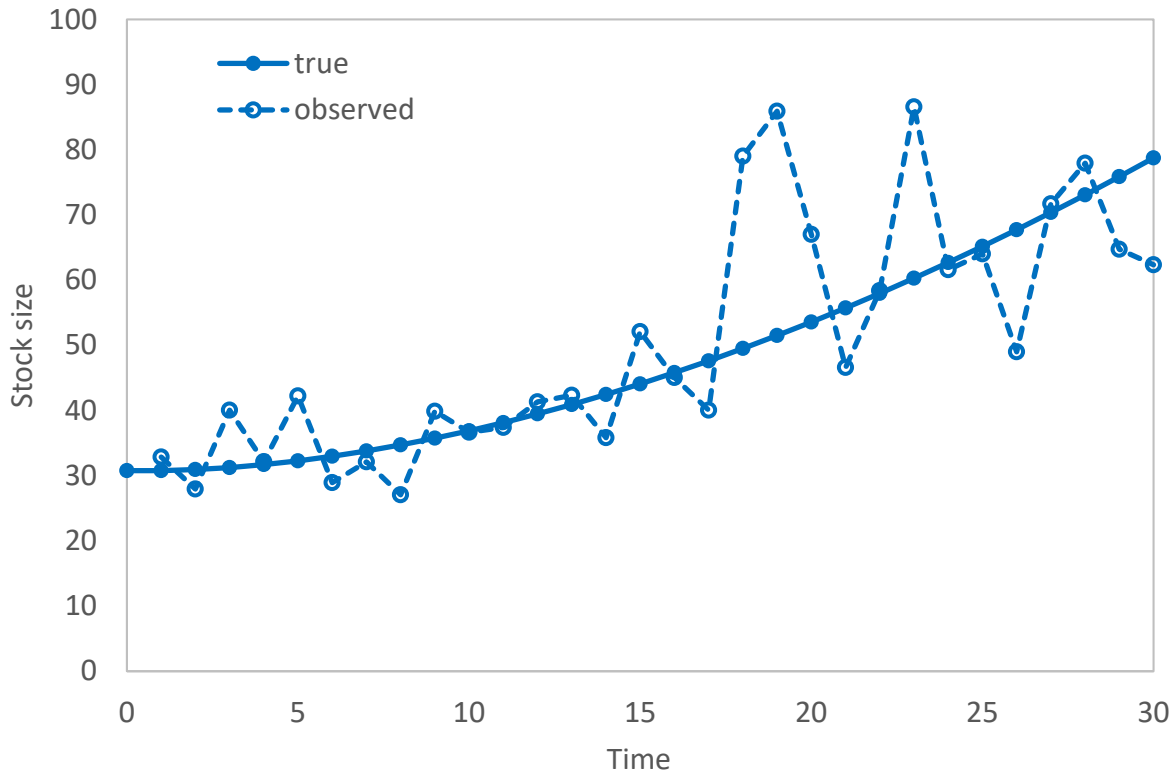

Figure A2.2. Time series of true and observed stock sizes.

Appendix 3

A general loss function

The loss function used in the main text can be viewed as a special case of the following:

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

where z is the quantity to be estimated, $\hat{z}$ is the estimator of z, and ra is a real number. Note that, in order for ra to be unrestricted, z must be restricted to non-negative values only. Setting $ra = 0$ gives the special case used in the main text. In the limit as ra approaches unity, the above loss function converges to $(\ln(z) - \ln(\hat{z}))^2$.

Note that this function does not exhibit constant risk aversion, either relative or absolute, in the sense of Pratt (1964) and Arrow (1965, 1971), although the parameter ra can be viewed as an *ad hoc* measure of risk aversion, so long as z exhibits the property that any *under*estimate is preferred to an *over*estimate of the same magnitude.

For probability density function $f(z)$, the risk (i.e., the expected loss) associated with $\hat{z}$ is given by

$$risk(\hat{z}|ra) = \int_0^\infty f(z)\, loss(z|\hat{z},ra)dz = \frac{m(2(1-ra))^{2(1-ra)} - 2m(1-ra)^{1-ra}\hat{z}^{1-ra} + \hat{z}^{2(1-ra)}}{(1-ra)^2}.$$

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

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

which is solved by setting $\hat{z} = m(1-ra)$. Some examples: if $ra = 0$, the optimal estimator is the arithmetic mean; if $ra = 1$, the optimal estimator is the geometric mean; and if $ra = 2$, the optimal estimator is the harmonic mean.

Given that the optimal estimator is an order mean, it is worthwhile noting that the order mean of the weighted average distribution is equal to the order mean of the individual (cross-conditional) order means. For example, the harmonic mean of the CCMA distribution is equal to the harmonic mean of the individual (cross-conditional) harmonic means. To see this, recall that the CCMA distribution is approximated as follows:

$$f_{ave}(\boldsymbol{\zeta}) = \sum_{i=1}^{nmod}\left(p_i \sum_{j=1}^{nmod} W_{i,j}\, h(Z_{i,j},\boldsymbol{\zeta})/nrep\right).$$

The order mean of the CCMA distribution is then given by

$$m_{ave}(q) = \left(\sum_{bin=1}^{nbin} f_{ave}(\boldsymbol{\zeta})_{bin}(\zeta_{bin})^q\right)^{1/q},$$

and the order means of the individual (cross-conditional) distributions are given by

$$M_{i,j}(q) = \left(\sum_{bin=1}^{nbin} \left(\frac{h\left(Z_{i,j}, \boldsymbol{\zeta}\right)_{bin}}{nrep} \right) (\zeta_{bin})^q \right)^{1/q} .$$

Combining the above gives:

$$m_{ave}(q) = \left(\sum_{i=1}^{nmod} \left(p_i \sum_{j=1}^{nmod} W_{i,j} \, M_{i,j}(q)^q \right) \right)^{1/q} .$$

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.