ICES Journal of Marine Science Advance Access published November 20, 2014

ICES Journal of Marine Science

ICES International Council for the Exploration of the Sea
CIEM Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science; doi:10.1093/icesjms/fsu198

Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models

Felipe Hurtado-Ferro[1]*, Cody S. Szuwalski[1,2], Juan L. Valero[3], Sean C. Anderson[4], Curry J. Cunningham[1], Kelli F. Johnson[1], Roberto Licandeo[5], Carey R. McGilliard[6‡], Cole C. Monnahan[7], Melissa L. Muradian[7], Kotaro Ono[1], Katyana A. Vert-Pre[8], Athol R. Whitten[1], and André E. Punt[1]

[1]*School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA, USA*
[2]*Bren School of Environmental Science and Management, University of California, Santa Barbara, CA 93106-5131, USA*
[3]*Center for the Advancement of Population Assessment Methodology, 8901 La Jolla Shores Drive, La Jolla, CA 92037, USA*
[4]*Earth to Ocean Research Group, Department of Biological Sciences, Simon Fraser University, Burnaby, BC, Canada V5A 1S6*
[5]*Fisheries Centre, Aquatic Ecosystems Research Laboratory, University of British Columbia, Vancouver, BC, Canada V6T 1Z4*
[6]*Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Alaska Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, Washington, United States of America*
[7]*Quantitative Ecology and Resource Management, University of Washington, Box 352182, Seattle, WA 98195-5020, USA*
[8]*School of Forest Resources and Conservation, University of Florida, Box 110240, Gainesville, FL 32611, USA*

**Corresponding author: e-mail: fhurtado@uw.edu*
[‡]*Present address: National Marine Fisheries Service, Alaska Fisheries Science Center, National Oceanic and Atmospheric Administration, 7600 Sand Point Way NE, Seattle, WA 98115, USA.*

Hurtado-Ferro, F., Szuwalski, C. S., Valero, J. L., Anderson, S. C., Cunningham, C. J., Johnson, K. F., Licandeo, R., McGilliard, C. R., Monnahan, C. C., Muradian, M. L., Ono, K., Vert-Pre, K. A., Whitten, A. R., and Punt, A. E. Looking in the rear-view mirror: bias and retrospective patterns in integrated, age-structured stock assessment models. – ICES Journal of Marine Science, doi: 10.1093/icesjms/fsu198.

Received 30 May 2014; revised 3 October 2014; accepted 16 October 2014.

Retrospective patterns are systematic changes in estimates of population size, or other assessment model-derived quantities, that occur as additional years of data are added to, or removed from, a stock assessment. These patterns are an insidious problem, and can lead to severe errors when providing management advice. Here, we use a simulation framework to show that temporal changes in selectivity, natural mortality, and growth can induce retrospective patterns in integrated, age-structured models. We explore the potential effects on retrospective patterns of catch history patterns, as well as model misspecification due to not accounting for time-varying biological parameters and selectivity. We show that non-zero values for Mohn's ρ (a common measure for retrospective patterns) can be generated even where there is no model misspecification, but the magnitude of Mohn's ρ tends to be lower when the model is not misspecified. The magnitude and sign of Mohn's ρ differed among life histories, with different life histories reacting differently from each type of temporal change. The value of Mohn's ρ is not related to either the sign or magnitude of bias in the estimate of terminal year biomass. We propose a rule of thumb for values of Mohn's ρ which can be used to determine whether a stock assessment shows a retrospective pattern.

Keywords: bias, fisheries stock assessment, integrated analysis, retrospective patterns, simulation, statistical age-structured models.

Introduction

A retrospective pattern (or bias) has been defined as a "systematic inconsistency among a series of estimates of population size, or related assessment variables, based on increasing periods of data" (Mohn, 1999; Figure 1 upper panels). Butterworth (1981) first described a retrospective pattern in estimated biomass of the Southwest African pilchard (*Sardinops ocellata*) stock, and Sinclair *et al.* (1991) showed similar patterns in Northwest Atlantic groundfish assessments. Since then, Mohn (1999), ICES (2002, 2003, 2004, 2007), NOAA (2009), and Deroba (2014) have examined the issue

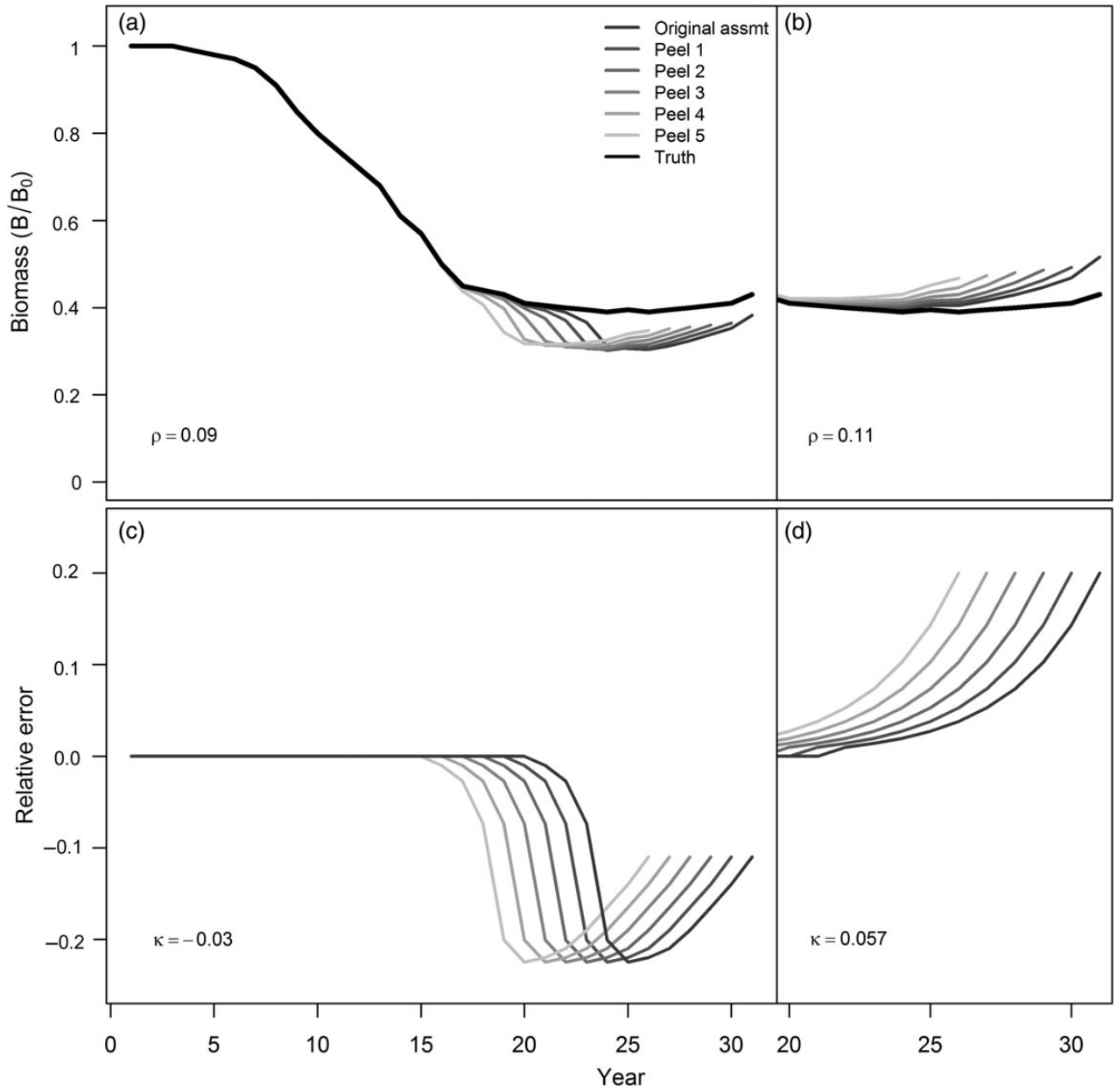

Figure 1. Two generated retrospective patterns (a and b) and their relative errors (c and d), along with the corresponding values for Mohn's ρ and κ-statistics.

more thoroughly using simulation. Although the possibility of retrospective patterns is now considered in many stock assessments and there are many case studies, few papers have investigated the origins of retrospective patterns (Cadigan and Farrell, 2005, being an exception).

Retrospective patterns in estimated biomass present large challenges for fisheries stock assessment and management. Total allowable catches in each year are generally based on an estimate of biomass and some agreed level of fishing mortality. A systematically biased estimate of biomass can result in not only recommended catches for a single year that are higher or lower than intended, but also for several consecutive years. From a conservation point of view, retrospective patterns are of particular concern when they lead to overestimation of the biomass used to set catch levels, whereas they could lead to underutilization if biomass is systemically underestimated. In the former cases, fishery scientists and managers may believe that there are more fish in the sea than appears based on subsequent assessments, and catches can be set at levels exceeding targets while being unnoticed for several years (e.g. Pacific halibut, *Hippoglossus stenolepis*; Valero, 2012). On the other hand, when a retrospective pattern is noticed, it can be so severe that an assessment could be considered unreliable for management purposes (Cadigan and Farrell, 2005; Valero, 2012).

The root causes of a specific retrospective pattern are often difficult to determine, given available data. Parma (1993) was the first to ascribe a potential causal mechanism (time-varying catchability) to an observed retrospective pattern. NOAA (2009) demonstrated a large number of ways in which retrospective patterns can be produced in simulated populations to which virtual population analysis (VPA) methods were applied to estimate biomass and other

quantities important in management. Generally, retrospective patterns arise from two general pathologies: time-varying processes unaccounted for in the assessment (i.e. model misspecification), or contradictory (or incomplete) data. To date, simulation studies have used primarily VPA-type stock assessment methods, but integrated, statistical age-structured models are also extensively used around the world (Maunder and Punt, 2013) and have also shown large retrospective patterns in some assessments (e.g. Pacific halibut *Hippoglossus stenolepis*; Valero, 2012; Norton sound red king crab *Paralithodes camtschaticus*; Hamazaki and Zheng, 2012). Nonetheless, it is not known how retrospective patterns emerge in integrated stock assessments, though model misspecification and conflicting data are likely culprits.

Measuring retrospective patterns has also proven challenging. The most commonly used metric is the "ρ" statistic proposed by Mohn (1999), which measures the relative difference between an estimated quantity from an assessment with a reduced time-series and the same quantity estimated from the full time-series. However, despite this being a relative measure, no rules of thumb have been developed with respect to how large the value of Mohn's ρ must be before an assessment is deemed to have a retrospective pattern (NOAA, 2009). It is also unclear how much information does Mohn's ρ provide about the bias in the final year of an assessment.

In this study, we aim to understand how model misspecification in three processes modelled in integrated age-structured models (natural mortality, growth, and selectivity) may generate retrospective patterns, and how these patterns vary across life histories and exploitation patterns. We propose ranges of Mohn's ρ that can be used as guidelines to determine whether an assessment exhibits retrospective patterns that are substantial enough to be of concern to assessment scientists and managers based on life history, and explore methods to determine the cause of retrospective patterns.

Methods

Overview

Monte Carlo simulations were implemented using the ss3sim simulation framework (Anderson *et al.*, 2014a, b), an open-source software package implemented in the R statistical software environment (R Core Team, 2014). ss3sim has been used for stock assessment simulation studies (Johnson *et al.*, 2014; Ono *et al.*, 2014), and is built around Stock Synthesis 3 (SS; Methot and Wetzel, 2013), a software platform using Automatic Differentiation Model Builder (ADMB; Fournier *et al.*, 2012). SS is used widely to implement integrated, statistical models for fisheries stock assessment (see Appendix B of Methot and Wetzel, 2013). The ss3sim simulation framework consists of three parts: a conditioning model (CM), an actual stock assessment model used to parameterize the operating model; an operating model (OM), which generates the "true" population dynamics of the system from which data are sampled; and a separate estimation model (EM), which is fit to the data and provides estimates of quantities important for management, and represents a manager's perception of the system (Figure 2). This simulation framework provides insight unobtainable from actual assessments because it allows for the direction and magnitude of retrospective patterns to be linked to the true changes in the population processes that caused them.

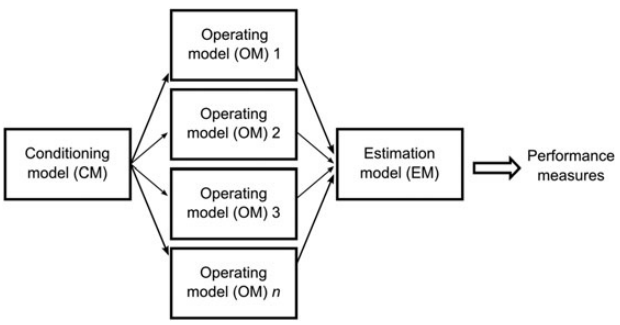

Figure 2. General design of the simulation study.

Model description

The OM and EM were both age-structured. Each simulation considered a 100-year period during which a single fishery operated starting in year 25 and a single survey operated every other year beginning in year 60 (Figure 3). Catch was reported annually without error from the start of the fishery to year 100 (terminal year). The survey index of abundance was generated with lognormal error ($CV = 0.2$). Fishery and survey length- and age-composition data were generated using a multinomial distribution with 100 samples, and the correct effective sample size was passed to the EM (see Anderson *et al.*, 2014a, b, for more detail). Fishery and survey selectivity were assumed (correctly) to be asymptotic in the EM. Recruitment dynamics were specified in the OM using a Beverton–Holt stock–recruitment relationship, and the EM assumed the correct functional form. All estimated parameters were assumed to be time invariant, although this was not the case in many of the OMs (Table 1). Thus, the EM was not misspecified except when some parameters were time-varying.

The EM estimates virgin recruitment (R_0), deviations in recruitment about the stock–recruitment curve, fishery and survey selectivity parameters, survey catchability (q), and somatic growth parameters (L_∞, K). Natural mortality (M), the steepness of the stock–recruitment relationship (h), and the extent of variation about the stock–recruitment relationship (σ_R) were assumed to be known without error.

Experimental design

Four factors were explored in this study: (i) life history type; (ii) time-varying process; (iii) fishing mortality (F) pattern; and (iv) how the time-varying process varied over time.

(i) Life history: Populations were simulated for three general life history patterns: North Sea cod ("Cod"; *Gadus morhua*; OM parameter values supplied by R. Methot, NMFS, NOAA, pers. comm.), yellowtail flounder ("Flatfish"; *Limanda ferruginea*; OM parameter values from Legault *et al.*, 2012), and Pacific sardine ("Sardine"; *Sardinops sagax caeruleus*; OM parameter values taken from Hill *et al.*, 2012). The Ricker stock–recruitment function used by Hill *et al.* (2012) for the sardine-like life history was replaced by a Beverton–Holt stock–recruitment function with a steepness specific to sardine (Myers *et al.*, 1999) to facilitate comparisons among the three life history types. Parameters used for each life history type are shown in Table 1.

(ii) Time-varying process: Growth, fishery selectivity, and M were allowed to vary over time (individually) in the OM in an attempt to induce retrospective patterns. Growth was changed so that the age at which 95% of individuals reach

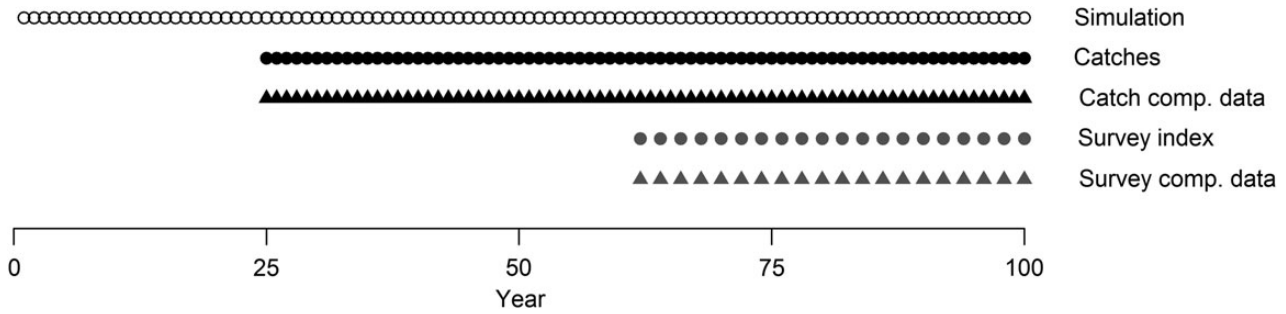

Figure 3. Extent of data available to the EM. Both catches and the survey have associated length- and age-composition data. The survey index was generated as lognormal samples with a *CV* of 0.2. The survey occurs every 2 years for 20 years.

Table 1. Life history, fishery, and modelling parameters used for each life history type (cod-like, flatfish-like, and sardine-like).

Parameter (units)	Symbol	Estimated	Cod	Flatfish	Sardine
Base parameters					
Natural mortality (year^{-1})	M	No	0.2	0.2	0.4
Reference age (year)	a_1	No	0.5	0.5	0.5
Maximum age (year)	A_{max}	No	25	25	15
Biology					
Length at a_1 (cm)	L_1	Yes	20	12.7	10
Length at A_{max} (cm)	L_∞	Yes	132	47.4	25
Growth rate (year^{-1})	K	Yes	0.2	0.35	0.4
CV L_1	CV_1	Yes	0.1	0.2	0.14
CV L_∞	CV_∞	Yes	0.1	0.2	0.05
Length-weight scaling (kg cm^{-3})	α	No	6.8e−6	1.0e−5	1.7e−5
Allometric factor	β	No	3.1	3.0	2.9
Maturity slope (cm^{-1})	Ω_1	No	−0.27	−0.42	−0.90
Length at 50% maturity (cm)	Ω_2	No	38.2	28.9	15.9
Recruitment					
Log mean virgin recruitment	ln R_0	Yes	18.7	10.5	16
Steepness	h	No	0.65	0.76	0.59
Recruitment variability	σ_r	No	0.4	0.7	0.73
Selectivity					
Mean fishery length-at-50% selectivity (cm)	S_1	Yes	38.2	28.9	15.9
Fishery length selectivity slope (cm)	S_2	Yes	10.6	7	3.3
Survey length-at-50% selectivity (cm)	S_3	Yes	30.5	23.1	12.7
Survey length selectivity slope (cm)	S_4	Yes	10.6	7	3.3
Log-catchability	ln q	Yes	0	0	0
Survey observation error s.d.	σ_S	No	0.2	0.2	0.2
Time-varying parameters (final values)					
K (increase)	–	No	0.2731	0.4871	0.548
K (decrease)	–	No	0.1578	0.2736	0.3149
S_1 (increase)	–	No	47.745	36.125	19.87
S_1 (decrease)	–	No	28.655	21.675	11.93
M (increase)	–	No	0.255	0.3	0.57
M (decrease)	–	No	0.165	0.14	0.29

L_∞ increased or decreased by 25% from the initial values (Figure 4a). Selectivity was changed such that the length at which 50% of individuals were selected increased or decreased by 25% from the initial values (Figure 4b). M was changed such that the new maximum sustainable yield (*MSY*) was $\pm$25% of the original *MSY* (Figure 4c). Table 1 shows base and modified values for these parameters.

(iii) Fishing mortality pattern: Three typical patterns in F were used to generate the catches: constant F, equal to the value that produced 0.95 *MSY* on the left limb of the yield vs. F curve (Figure 4d); a steadily increasing trend to the F corresponding to 0.95 *MSY* on the right limb of the yield vs. F curve (Figure 4e); and a "fish down and recovery", i.e. a 60-year linear increase to the F corresponding to 0.95 *MSY* (right limb), followed by a 15-year linear decrease to the F corresponding to 0.95 *MSY* (left limb; Figure 4f). For all scenarios, years 1 through 25 had zero fishing, and acted as a burn-in period.

(iv) Patterns in time-varying processes: The onset of the change in a process occurred either 10 years ("Recent") or 25 years ("Old") before the last year of the simulation; the pattern of the change was either "sudden" or "gradual"; and the direction

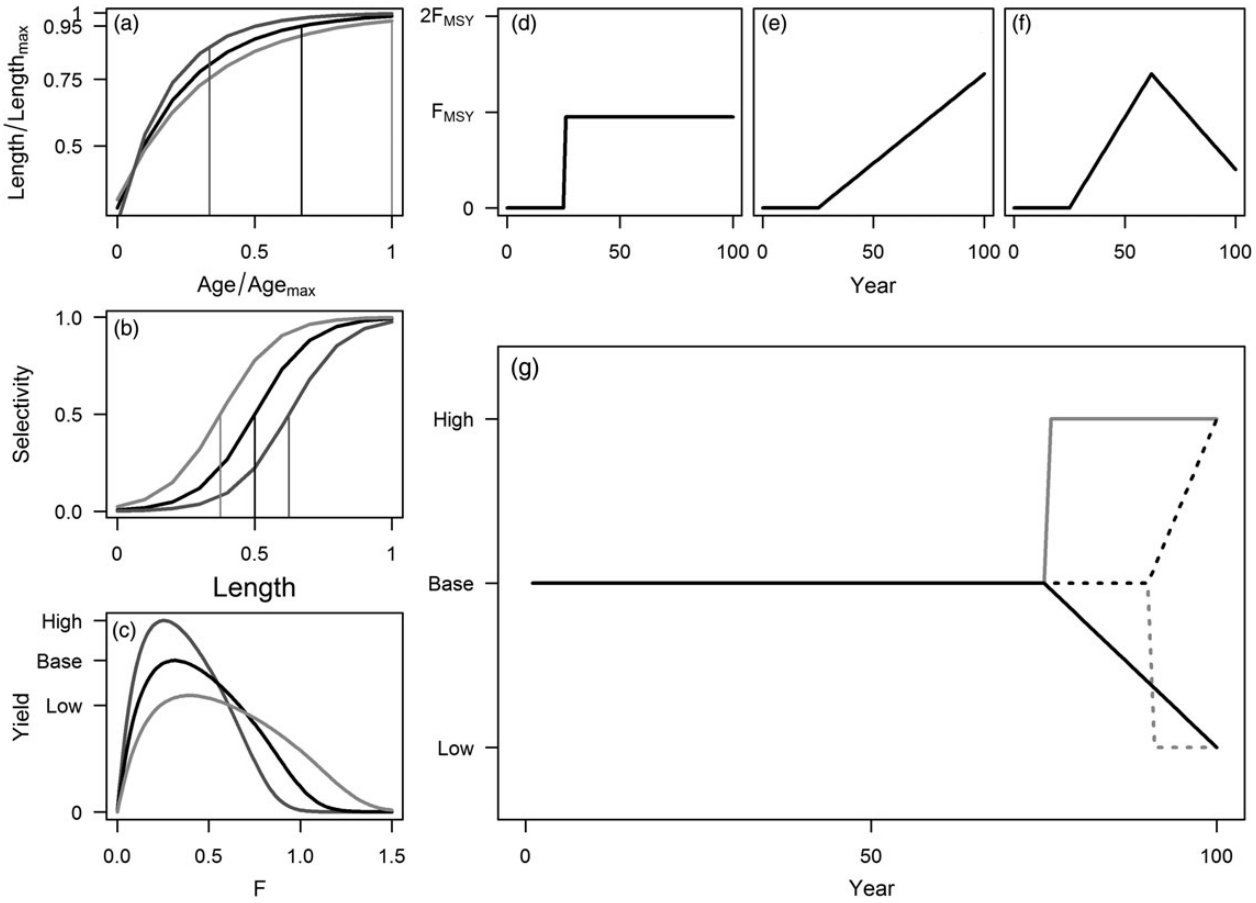

Figure 4. Experimental design showing the time-varying processes (a–c), fishing patterns (d–f), and time-varying patterns for the processes (g). For the patterns of time variance, only some cases are shown to reduce clutter (solid and dotted lines show old and recent timing, respectively; black and grey lines show gradual and sudden patterns, respectively). See text for further explanation.

Table 2. Scenarios explored in this paper.

Scenario	Timing	Direction	Pattern
Base	None	None	None
1	Old	Positive	Gradual
2	Recent	Positive	Gradual
3	Old	Negative	Gradual
4	Recent	Negative	Gradual
5	Old	Positive	Sudden
6	Recent	Positive	Sudden
7	Old	Negative	Sudden
8	Recent	Negative	Sudden

of the changes, "positive" and "negative", were tested for the ability to induce retrospective patterns (Figure 4g). Table 2 details the eight time-varying scenarios.

A full factorial design with the 216 possible combinations of all four factors was performed. Additionally, nine "base" cases (three life histories × three fishing patterns) with no time-varying processes were also evaluated for comparison. Fifty simulations were performed for each scenario. The EM was run six times for each of the 50 simulations: once with all available data and five times with one fewer year of data each time (model runs with fewer data are referred to as "peels").

Convergence criteria

Convergence was determined using the maximum gradient from the minimization procedure. Only simulations with gradients <0.1 were accepted. To ensure that all scenarios had the same number of simulations, new iterations using new randomly generated numbers were run until the desired number of simulations was obtained. Given time constraints, inverting the Hessian matrix (a common test of convergence) was not an option.

Performance metrics and analyses

Several performance metrics were used. Mohn's ρ was calculated for estimated biomass, recruitment, and F. Mohn's ρ is defined as:

$$\rho = \overline{\left(\frac{X_{Y-y,p} - X_{Y-y,\mathrm{ref}}}{X_{Y-y,\mathrm{ref}}}\right)}, \quad (1)$$

where X is the quantity for which Mohn's ρ is being calculated, Y the final year of the simulation, y the last year of a given "peel" p, and ref the reference peel, i.e. the most recent assessment. Note that this formulation is slightly different from that given by Mohn (1999), where instead of summing across peels, these are averaged (R. Mohn, pers. comm.).

An index (κ) was developed to determine whether the biomass trajectories converge towards or diverge away from the true

biomass (termed "convergent" or "divergent" retrospective patterns; see Figure 1). A divergent pattern (positive κ) means that the average absolute bias in the last year of a peel is larger than the absolute bias in the previous to last year of the peel. This index is defined as:

$$\kappa = \frac{\sum_{p=1}^{n} |RE_{Y-p,p}| - |RE_{Y-p-1,p}|}{n}, \quad (2)$$

where Y is the final year of the simulation, p the number of years being "peeled", n the number of "peels" performed, and RE the relative error

$$RE_{y,p} = \frac{\hat{X}_{y,p} - X^{\text{true}}_{y,p}}{X^{\text{true}}_{y,p}}, \quad (3)$$

for quantity X (where $\hat{X}$ is the estimate of X and X^{true} is the true value of X), peel p, in year y. The κ index differs from Mohn's ρ in that it compares errors against the true value of X, rather than against the most recent assessment. The index κ could be used for any quantity of interest, like recruitment or F time-series.

To compare different scenarios quantitatively, a fixed-effects analysis of variance (ANOVA) was used to evaluate the proportion of the variance of the values of Mohn's ρ for spawning biomass explained by each factor. Separate analyses were conducted by life history type because differences among life histories are large and may dampen the effect of individual factors. Values for Mohn's ρ were also regressed against the relative error of biomass in the terminal year from the last (most data-rich) assessment to determine whether Mohn's ρ gives any information about the bias in model estimates. Note that these analyses were not conducted to evaluate "statistical significance", but rather as a way to characterize the output of a complicated set of simulations.

Non-metric multidimensional scaling (NMDS; Kurskal, 1964) was used to characterize the influence of specific time-varying processes in the OM not captured in the EM on parameter estimates. Relative errors in estimated parameters and derived quantities (e.g. biomass) were used as response variables; the factors in the experimental design were the predictor variables. NMDS is a tool for distilling information from many potentially correlated response variables. NMDS was used here because it does not require the assumption of multivariate normality and was chosen for this dataset because of potential non-linearities. NMDS avoids the assumption of normality by ordinating objects in low-dimensional space using ranks of a dissimilarity/distance matrix, rather than raw values. It is an iterative procedure in which an algorithm attempts to minimize the "stress" between the ordination and the ranked distance matrix (see Kurskal, 1964, for details). Stress is a measure of the "optimality" of an ordination (smaller is better) and is defined as:

$$\text{stress} = \sqrt{\frac{\sum_i \sum_j (d_{i,j} - \tilde{d}_{i,j})^2}{\sum_i \sum_j d_{i,j}}}, \quad (4)$$

where $d_{i,j}$ is the original distance in the full dimensional space, $\tilde{d}_{i,j}$ the adjusted distance in the ordination of the chosen number of dimensions, and i and j are the indices of the dissimilarity matrix. The dissimilarity/distance matrix was calculated using Euclidean distance, after standardizing the data by subtracting the mean and dividing by the standard deviation for each covariate. A correlation analysis between covariates associated with each trial and the scores from each NMDS axis were used to examine relationships. A permutation analysis ($n = 1000$) was used to assess the significance of the relationships between the covariates and the performance statistics.

Results

Retrospective patterns were found under all three processes and factors evaluated (Figure 5). The magnitude and direction of these patterns varied among processes and life histories, but several general patterns emerged. The base cases for the three life history types and F patterns were essentially unbiased (median Mohn's ρ was 0.01, −0.01, and −0.02 for cod, flatfish, and sardine,

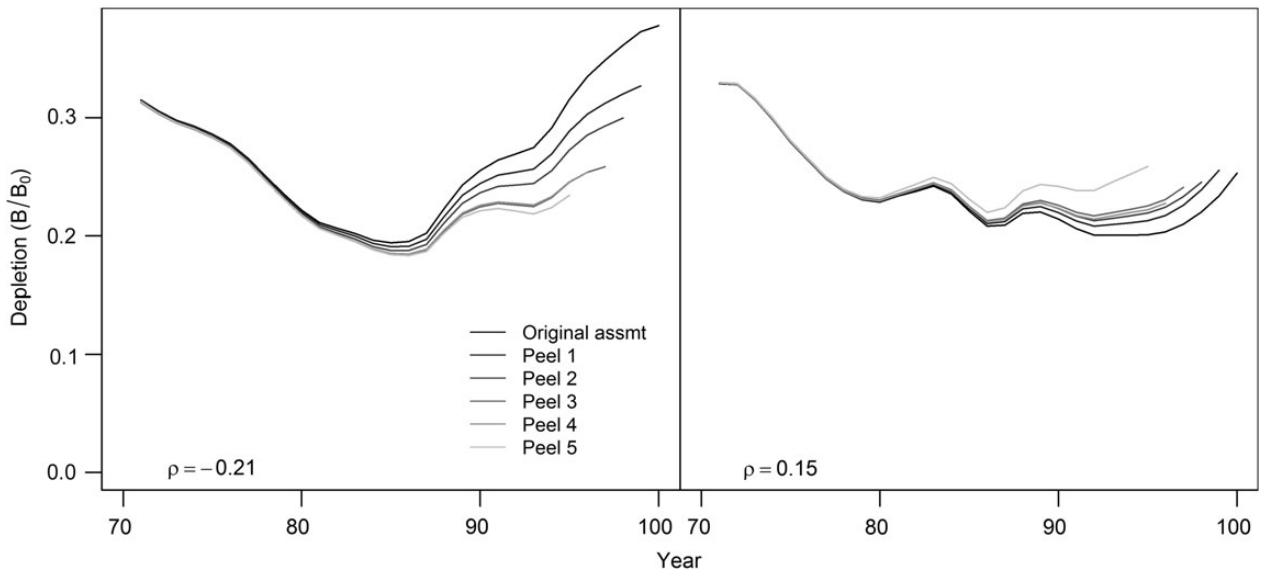

Figure 5. Sample results showing retrospective patterns for cod for fishing mortality "fish down and recovery". Growth is time varying in the OM, with results for scenario 2 (a; recent, negative, gradual change) and scenario 3 (b; old, positive, gradual change).

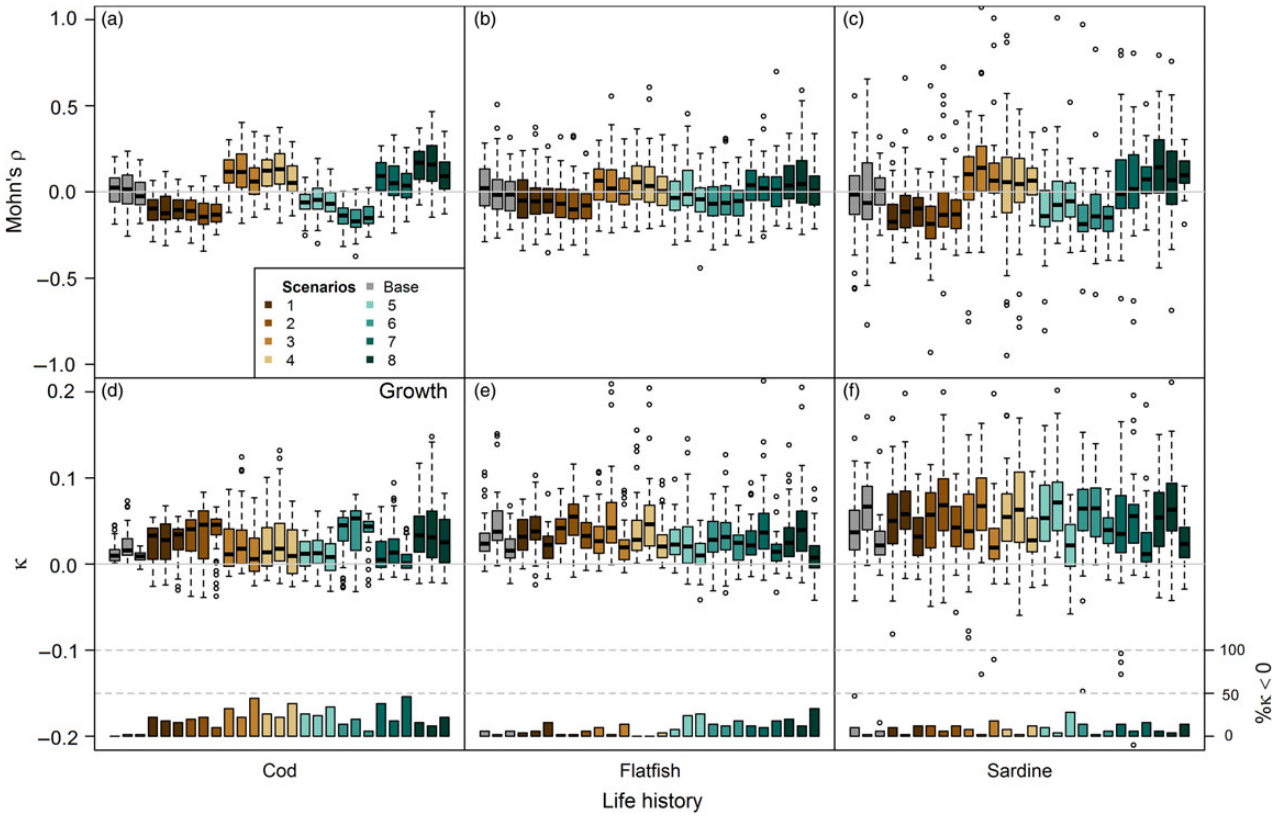

Figure 6. Distribution of Mohn's ρ (upper panels) and κ (lower panels) for spawning biomass, for cod (a and d), flatfish (b and e), and sardine (c and f) when growth is time varying. The bars on the lower *x*-axis show the percentage of times κ was <0. Colours are repeated three times for the three *F* patterns (constant, increasing, and up and down, respectively). See Table 2 for scenario IDs.

respectively), but still produced non-zero values for Mohn's ρ, indicating that pure random noise can produce what appear retrospective patterns. This is not surprising because the positive and negative peels will not cancel out (except by chance). However, the magnitude of Mohn's ρ for these cases was generally lower than those arising from model misspecification ("base" in Figures 6–8). Values for Mohn's ρ varied the most among replicate simulations for sardine, while cod was the least variable in this respect.

Effect of time-varying factors on Mohn's ρ statistic

For all life history types, a change to faster growth, selection at larger sizes, and lower *M* resulted in retrospective patterns with negative Mohn's ρ for estimates of biomass, while a change to slower growth, selection at smaller sizes, and higher *M* resulted in positive Mohn's ρ values (Figures 6–8). A positive value for the ρ statistic means that the quantity being evaluated is consistently being overestimated (when compared with the estimate from the full time-series) and is potentially most problematic in terms of sustainability. Fishing pattern showed a minor influence on the relationship between Mohn's ρ on the how time-varying growth impacted estimation performance (Table 3). Timing and pattern (i.e. abrupt vs. gradual) of the changes did not have a noteworthy influence.

Other, more complex, patterns also emerged, and some factors may be more important for some life history types than others (Table 3). Comparisons of which variable (*M*, selectivity, growth) had the largest impact on Mohn's ρ should be interpreted with caution because the changes over time in these quantities are not necessarily comparable. Nevertheless, for cod, the magnitude of Mohn's ρ was more marked for changes in growth (Figure 6a), but smaller for changes in selectivity (Figure 7a). Flatfish showed the opposite pattern, larger impacts for changes in selectivity (Figure 7b), but smaller for changes in growth (Figure 6b). All three life histories showed largest (positive or negative) Mohn's ρ values when natural mortality changed (Figure 8a–c); sardine showed values as large as those under changing natural mortality for all three factors studied (Figures 6c, 7c, and 8c).

Retrospective patterns for estimates of *F* showed a similar pattern as those for estimates of spawning biomass, but with different sign (Supplementary Figures S1–S3; Supplementary Table S1). However, the relationship between Mohn's ρ for biomass and that for *F* was not linear (Supplementary Figure S4). The curvature in the relationship (Supplementary Figure S4) is explained by the log-normal distribution of biomass, and the exponential relationship between biomass and *F*.

Convergence/divergence and bias

A positive value of the convergence/divergence index, κ, means that a retrospective pattern is divergent, i.e. the average absolute bias in the last year of a peel is larger than the absolute bias in the previous to last year of the peel (Figure 1d). The κ statistic was mostly positive for growth (Figure 6d and e) and natural mortality (Figure 8d and e) for the cod and flatfish life histories, and always for sardine (Figures 6f, 7f, and 8f). However, cod and flatfish showed negative values for κ about half of the time for selectivity (Figure 7d and e). The index κ showed a similar behaviour for retrospective patterns in *F* as it did for spawning biomass.

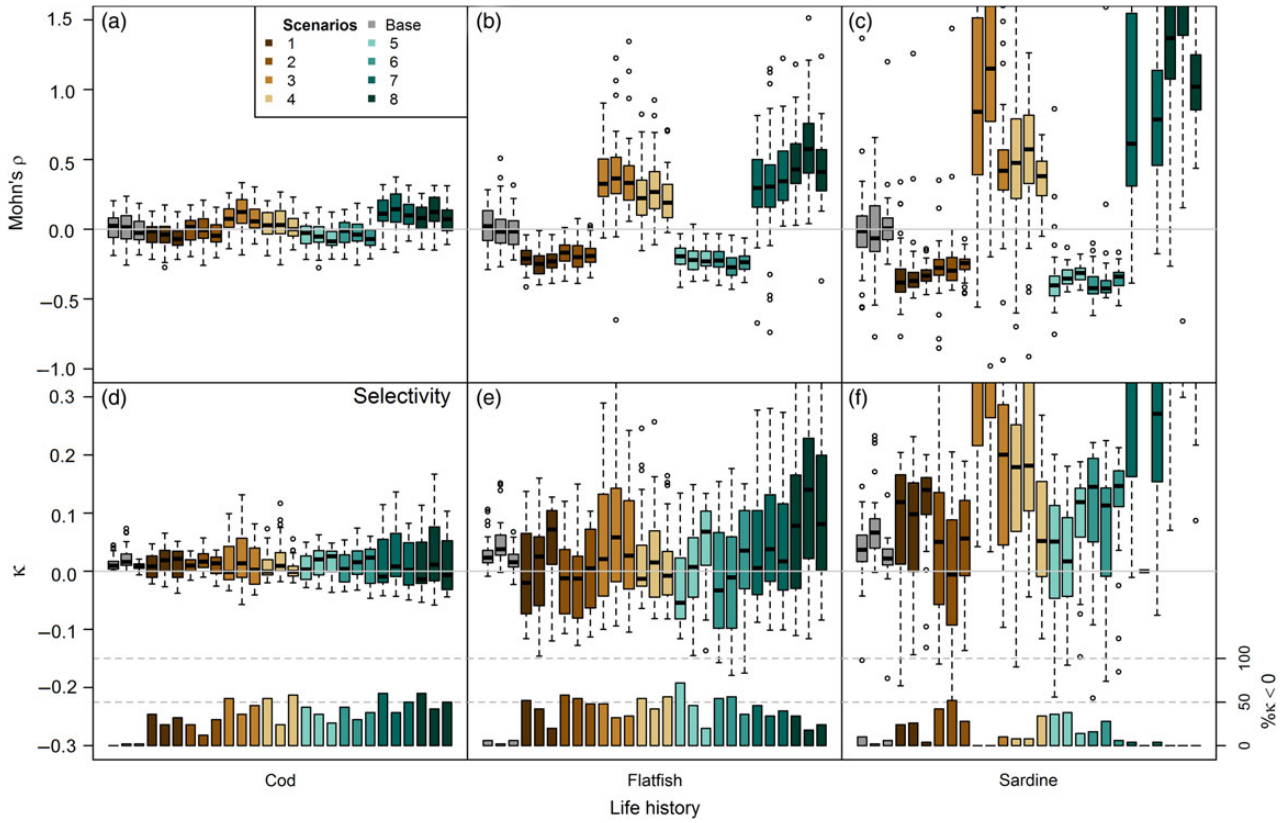

Figure 7. Distribution of Mohn's ρ (upper panels) and κ (lower panels) for spawning biomass, for cod (a and d), flatfish (b and e), and sardine (c and f) when selectivity is time varying. The bars on the lower *x*-axis show the percentage of times κ was <0. Colours are repeated three times for the three *F* patterns (constant, increasing, and up and down, respectively). See Table 2 for scenario IDs.

The value of Mohn's ρ was not related to bias in biomass in the terminal year of the full-data assessment, indicating that neither the magnitude nor the sign of Mohn's ρ provides any information about how biased an assessment might be (Figure 9).

Influence of specific time-varying processes in the OM on estimates of parameters

NMDS was only performed for the flatfish and cod life histories because of the large variability in the response of sardine-like species to the processes tested (Figure 10). The final stress level of the ordination was 11.2%, which indicates an informative ordination of the variability of the relative errors in parameters and measures of retrospective bias in biomass and *F*. All the responses (i.e. the relative errors and biases) were significant ($p < 0.001$), but had varying goodness-of-fits. A relatively strong gradient in Mohn's ρ and κ exist for biomass ($r^2 = 0.45$ and 0.40, respectively) and the relative error for parameters related to growth. Gradients were small for the relative error in selectivity parameters (average $r^2 \sim 0.03$) and the log of R_0 (Supplementary Table S2).

Correlations between scenarios and the ordination scores for the responses were significant, but effect sizes (i.e. goodness-of-fits) were very small, which suggests the observed correlations may be an artefact of the large sample size resulting from a repeated simulation study. Still, some generalizations might be drawn from the NMDS. Figure 10 shows the scenarios and responses for which the *p*-values for the goodness-of-fit statistics were <0.01. Larger Mohn's ρ and κ for biomass were more associated with a steadily increasing *F* pattern and gradual changes in population processes. Larger Mohn's ρ for *F* and error in biomass were associated with sudden, recent changes in population processes and were related to larger relative errors in selectivity parameters in both the fishery and the survey. Relative errors in growth parameters are not all shown, but have similar gradients and directions as *K* and were associated with a "fish down with recovery" *F* pattern with recent changes in population processes.

Discussion

We have shown that retrospective patterns can be induced in integrated, age-structured models, when time-varying processes are not accounted for in an assessment. Model misspecification had been identified in the past as a cause of retrospective patterns (Mohn, 1999; Cadigan and Farrell, 2005; NOAA, 2009), but current literature is still unclear about how large a retrospective pattern has to be for it to be of concern. It is generally assumed that retrospective patterns cannot be generated from purely random processes (NOAA, 2009). Although Mohn's ρ for simulated data without underlying time-varying processes was roughly unbiased, convincing-looking retrospective patterns were generated in some cases from these data (Supplementary Figure S5). Such patterns tended to have smaller Mohn's ρ statistics than those produced by model misspecification, so Mohn's ρ should always be used (instead of visual identification) for identifying retrospective patterns that warrant corrective measures such as allowing for time-varying parameters in the stock assessment.

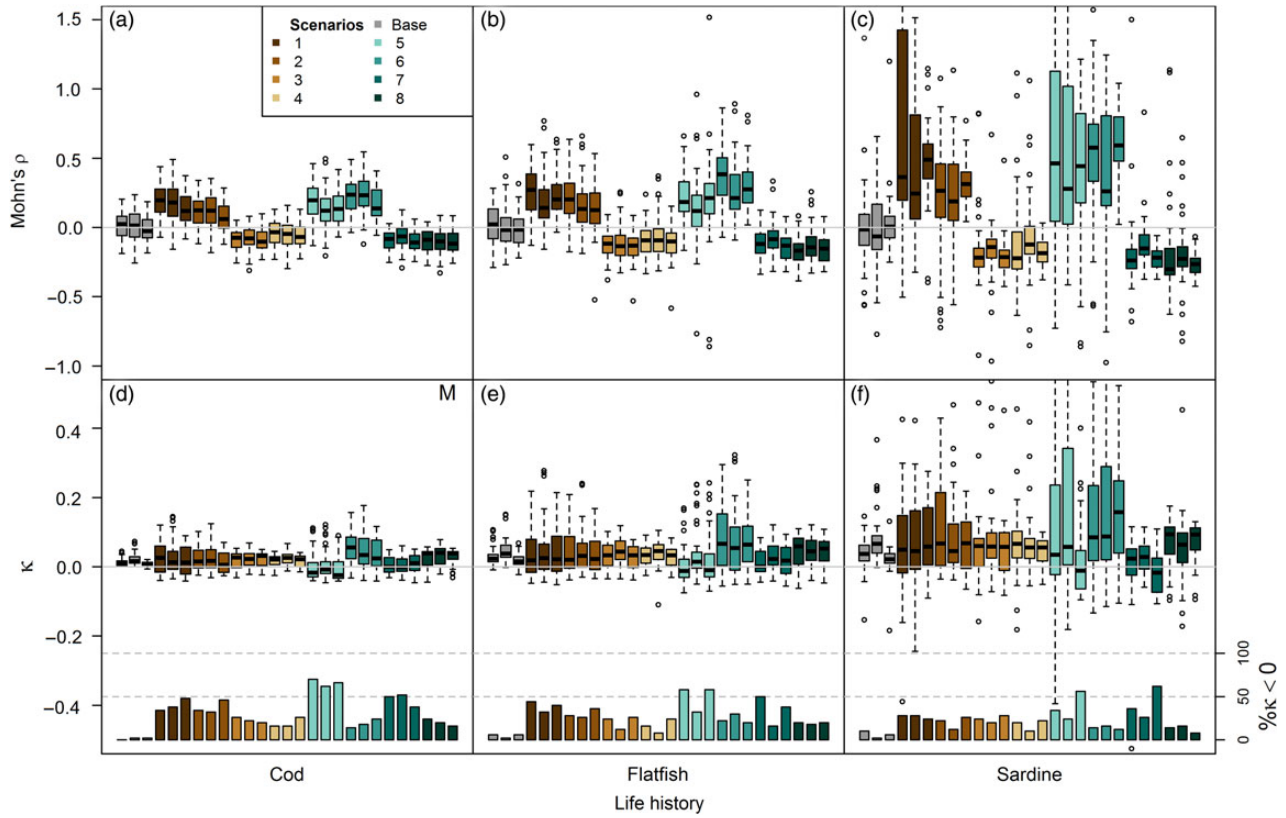

Figure 8. Distribution of Mohn's ρ (upper panels) and κ (lower panels) for spawning biomass, for cod (a and d), flatfish (b and e), and sardine (c and f) when M is time varying. The bars on the lower x-axis show the percentage of times κ was <0. Colours are repeated three times for the three F patterns (constant, increasing, and up and down, respectively). See Table 2 for scenario IDs.

Table 3. ANOVA results for Mohn's ρ for biomass, showing the percentage of variance explained by each variable (values are sum of sq./total sum of sq.).

Variable	Cod	Flatfish	Sardine
Process	2.799	2.648	1.879
F	0.527	0.114	0.413
Timing	0.014	0.000	0.030
Direction	0.163	0.376	0.031
Pattern	0.032	0.034	0.086
Process × timing	0.001	0.002	0.027
Process × direction	5.065	2.464	0.066
Process × pattern	0.517	0.181	0.137
Residuals	90.881	94.18	97.331

All the processes considered led to retrospective patterns, even if certain combinations of processes did not produce them. Because we explored processes related to biology, fishery, and population dynamics (growth rate, selectivity, and natural mortality, respectively), this suggests that retrospective patterns can arise from misspecification in any parameter of a stock assessment failing to include time variance, when that process is actually time-varying. Although the absence of a retrospective pattern may suggest that a stock assessment is correctly specified, this must not be taken as definitive confirmation (in a few cases, misspecified models did not show retrospective patterns).

Retrospective patterns and the magnitude of Mohn's ρ are life history dependent, and could be case-specific (Mohn, 1999; NOAA, 2009). We found that species with higher variability in their dynamics showed higher variability and magnitudes in Mohn's ρ, but the general behaviour of this statistic was robust to life history type. Retrospective patterns did not change considerably with fishing history or depletion, suggesting that these factors are not important for the magnitude or sign of the Mohn's ρ statistic. However, we only explored catch series that were perfectly known; imperfectly known catch histories can also generate retrospective patterns (Parma, 1993; Mohn, 1999; NOAA, 2009), and should be considered as an option when attempting to identify responsible processes behind retrospective biases. Larger changes in time-varying factors may also result in larger values of Mohn's ρ, as has been observed in some New England assessments (Chris Legault, NMFS, NOAA, pers. comm.), but this was not addressed directly in this study.

Given that the variability of Mohn's ρ depends on life history, and that the statistic appears insensitive to F, we propose the following rule of thumb when determining whether a retrospective pattern should be addressed explicitly: values of Mohn's ρ higher than 0.20 or lower than -0.15 for longer-lived species (upper and lower bounds of the 90% simulation intervals for the flatfish base case), or higher than 0.30 or lower than -0.22 for shorter-lived species (upper and lower bounds of the 90% simulation intervals for the sardine base case) should be cause for concern and taken as indicators of retrospective patterns. However, Mohn's ρ values smaller than those proposed should not be taken as confirmation that a given assessment does not present a retrospective pattern, and the choice of 90% means that a "false positive" will arise 10%

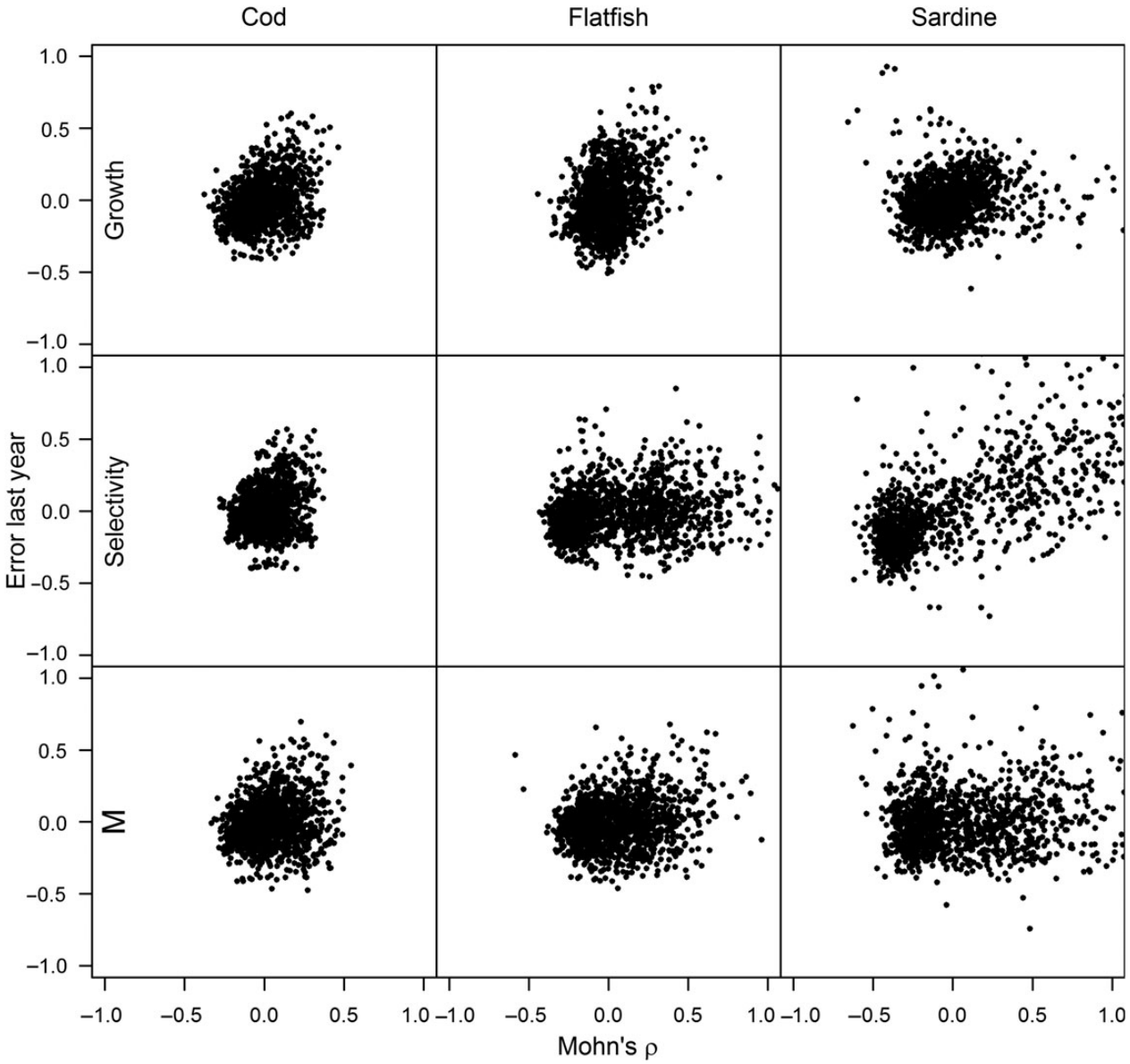

Figure 9. Relationship between Mohn's ρ and relative error in spawning biomass in the terminal year from the last (most data-rich) assessment.

of the time. In both cases, model misspecification would be correctly detected more than half the time.

This rule of thumb is presented with several caveats. The survey *CV* used in these simulations is small (0.2), which may affect the magnitude of retrospective patterns. We ran additional simulations for a subset of scenarios (cod with time-varying growth and constant harvest), with increasing survey uncertainty (*CV* 0.2, 0.4, 0.6, and 0.8) and estimating the hessian. Increasing survey uncertainty increased the variance of Mohn's ρ, but not its median (Supplementary Figure S6a). This increase in variance means that a "false positive" would arise 30% of the time when survey *CV* is 0.8. NOAA (2009) proposes a rule of thumb that considers estimation uncertainty to assess whether a retrospective pattern is large enough. Following this suggestion, we divided Mohn's ρ by the σ of the terminal estimate of biomass. This stabilized the variance of the index, but it also resulted in a decreasing absolute value of its median as survey uncertainty increased (Supplementary Figure S6b). This is undesirable, thus we consider that using the actual value of Mohn's ρ to be more appropriate for a rule of thumb. Also, we only used simulations from three representative life histories, and these intervals could vary for other life history types, such as large pelagic fish (e.g. tunas) or slow-growing, long-lived demersal species (e.g. rockfish).

The insensitivity of Mohn's ρ to the process which is varying over time (as seen through the small effect sizes of the different scenarios in the NMDS and the ANOVA) makes it a reliable statistic to assess whether retrospective patterns arise from time-varying processes. Mohn's ρ is useful in determining the direction of the change of the true time-varying process that may be causing the retrospective pattern. However, the magnitude of a Mohn's ρ is not related to bias in biomass or *F*, and should not be used to assess how far an assessment is from the truth.

For management purposes, some retrospective patterns are more concerning than others. Patterns that show a positive Mohn's ρ and

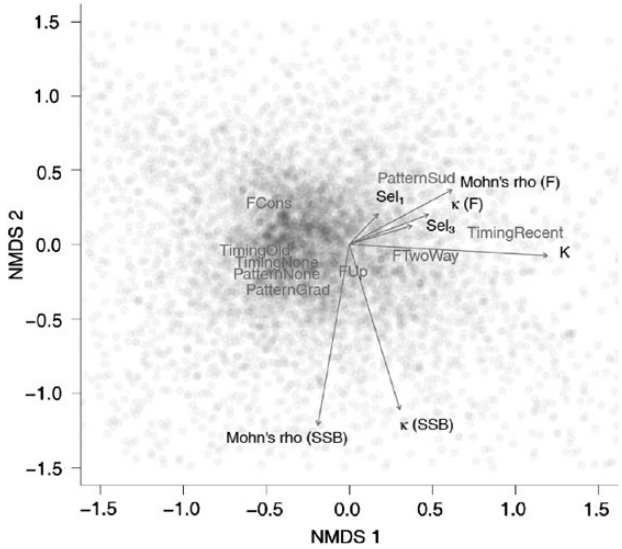

Figure 10. Ordination of simulation results via NMDS for cod and flatfish. Grey dots are individual simulations; arrows indicate gradients of significant (i.e. permutation *p*-values <0.01) response variables (relative error and measures of bias). The length of the arrow reflects the strength of the gradient. Grey text indicates the location of factors changed in scenarios that had permutation *p*-values <0.01. "FCons", "FUp", and "FTwoWay" represent constant *F*, increasing *F*, and fish down and recovery *F* patterns, respectively.

positive κ for biomass (negative Mohn's ρ for F) are the most concerning in terms of stock conservation, as they imply consistent overestimation of biomass and the highest risk for overfishing. Of the cases studied here, managers should be particularly alert to changes to slower growth or higher M, which resulted in the most positive Mohn's ρ and consequently, overestimation of biomass.

We showed that retrospective patterns arising from time-varying processes can look very similar and that NMDS was unable to associate any population process with gradients in bias statistics or relative error in estimated parameters. However, retrospective patterns in F are associated with higher relative errors in selectivity parameters. Retrospective patterns in F may be cause to examine selectivity more closely, for models with similar formulations to the ones included in our analysis. When retrospective patterns are observed in a stock assessment, they are often corrected by introducing estimation of a time-varying parameter (usually selectivity, M or q; Fu *et al.*, 2001; Legault *et al.*, 2011, 2012; Martell and Stewart, 2013) or applying a retrospective bias adjustment (TRAC, 2012; Deroba, 2014). The underlying causes for such patterns are commonly unknown (Fu *et al.*, 2001; ICES, 2008), but moving window analyses have proven promising in identifying the timing of a change which leads to the retrospective pattern (ICES, 2008). The risk of introducing further misspecification in an assessment can be high if these corrections are done based only on the presence of a retrospective pattern. It is uncertain whether a misspecification of time-varying parameters used to correct a retrospective pattern introduces more biases into the assessment. In addition, adding time-varying parameters could lead to an over-parameterized models and overall to poorer performance, so further research in this area is needed.

This study is a first attempt to systematically characterize retrospective patterns, and has caveats that should be considered. The κ-statistic is not available for real-life stock assessments, but we believe that it can be useful for further simulation studies. This statistic can be of particular interest when developing management strategy evaluations that consider retrospective patterns, as it informs whether these patterns are convergent or divergent. Whether a retrospective pattern is convergent or divergent has important implications for the conservation of the resource, as was discussed earlier. We studied a limited range of variables, and the behaviour of retrospective patterns to changes in other parameters (e.g. q, L_∞) should be explored systematically. Also, we only changed one process at a time, and did not explore the interaction between multiple changes in different parameters, or more complex patterns of time variance such as the pulse changes in catchability explored by NOAA (2009). We did not explicitly search for methods to identify the source of a retrospective pattern. Last, we only explored the effects of retrospective patterns on stock assessments, but did not evaluate how much risk they would introduce when managing a stock. For example, risk would depend on the status of the stock. If stock biomass is very high, a retrospective pattern might not be as problematic as it would be if biomass is very low, where a retrospective pattern can be more risky. Alternatively, model misspecification could affect the setting of reference points used in management. Studies such as management strategy evaluations (Smith *et al.*, 1999) are thus needed if these risks are to be characterized.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

The authors thank Chris Legault and one anonymous reviewer for their invaluable comments and suggestions. CSS was supported by a Washington SeaGrant fellowship. RRL was supported by Conicyt. MLM was funded by Exxon Valdez Oil Spill Trustee Council, grant 13120111-Q. AEP, KFJ, KO, CCM, and CRM were partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreements NA10OAR4320148, Contribution No. 2194, respectively. KFJ was partially supported for this work under a World Conference on Stock Assessment Methods travel bursary. SCA was supported by Fulbright Canada and NSERC. Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, R24 HD042828, to the Center for Studies in Demography and Ecology at the University of Washington. This research addresses the good practices in stock assessment modelling program of the Center for the Advancement of Population Assessment Methodology (CAPAM).

References

Anderson S. C., Monnahan C. C., Johnson K. F., Ono K., and Valero J. L. 2014a. ss3sim: an R package for fisheries stock assessment simulation with Stock Synthesis. PLoS ONE, 9: e92725.

Anderson S. C., Monnahan C. C., Johnson K. F., Ono K., Valero J. L., Cunningham C. J., Hurtado-Ferro F., *et al.* 2014b. ss3sim: fisheries stock assessment simulation testing with stock synthesis. R package version 0.8.0. http://cran.r-project.org/package=ss3sim.

Butterworth D. S. 1981. The value of catch-statistics-based management techniques for heavily fished pelagic stocks with special reference to the recent decline of the southwest African pilchard stock. *In* Applied

Operations Research in Fishing, pp. 441–464. Ed. by Haley . Springer, New York, US.

Cadigan N. G., and Farrell P. J. 2005. Local influence diagnostics for the retrospective problem in sequential population analysis. ICES Journal of Marine Science, 62: 256–265.

Deroba J. J. 2014. Evaluating the consequences of adjusting fish stock assessment estimates of biomass for retrospective patterns using Mohn's Rho. North American Journal of Fisheries Management, 34: 380–390.

Fournier D. A., Skaug H. J., Ancheta J., Ianelli J., Magnusson A., Maunder M. N., Nielsen A., *et al.* 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear modls. Optimization Methods and Software, 27: 233–249.

Fu C., Mohn R., and Fanning L. P. 2001. Why the Atlantic cod (*Gadus morhua*) stock off eastern Nova Scotia has not recovered. Canadian Journal of Fisheries and Aquatic Sciences, 58: 1613–1623.

Hamazaki T., and Zheng J. 2012. Norton sound red king crab stock assessment for the fishing year 2012/13. Stock Assessment and Fishery Evaluation Report for the King and Tanner Crab Fisheries of the Bering Sea and Aleutian Islands Regions, pp. 533–641. North Pacific Fishery Management Council, Anchorage, AK, USA.

Hill K. T., Crone P. R., Lo N. C. H., Demer D. A., Zwolinski J. P., and Macewicz B. J. 2012. Assessment of the Pacific sardine resource in 2012 for U.S. management in 2013. Pacific Fishery Management Council, Portland, OR.

ICES. 2002. Report of the Working Group on Methods on Fish Stock Assessments, ICES Headquarters, Copenhagen, Denmark, 3–7 December 2001. ICES CM 2002/D: 01.

ICES. 2003. Report of the Working Group on Methods on Fish Stock Assessments, ICES Headquarters, Copenhagen, Denmark, 29 January–5 February 2003. ICES CM 2003/D: 03.

ICES. 2004. Report of the Working Group on Methods on Fish Stock Assessments, Lisbon, Portugal, 11–18 February 2004. ICES CM 2004/D: 03.

ICES. 2007. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 13–22 March 2007, Woods Hole, USA. ICES CM 2007/RMC: 04.

ICES. 2008. Report of the Working Group on Methods of Fish Stock Assessments (WGMG), 7–16 October 2008, Woods Hole, USA. ICES CM 2008/RMC: 03.

Johnson K. F., Monnahan C. C., McGilliard C. R., Vert-pre K. A., Anderson S. C., Cunningham C. J., Hurtado-Ferro F., *et al.* 2014. Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. ICES Journal of Marine Science, doi:10.1093/icesjms/fsu055.

Kurskal J. B. 1964. Non metric multidimensional scaling: a numerical method. Psychometrika, 29: 115–129.

Legault C. M., Alade L., and Stone H. H. 2011. Stock Assessment of Georges Bank Yellowtail Flounder for 2011. Transboundary Resources Assessment Committee. http://www.nefmc.org/tech/cte_mtg_docs/110914/GF%20Docs/5.%20TRD_2011_01_GB_YTF.pdf.

Legault C. M., Alade L., Stone H. H., and Gross W. E. 2012. Stock Assessment of Georges Bank Yellowtail Flounder for 2012. Transboundary Resources Assessment Committee. http://www.bio.gc.ca/info/intercol/trac-cert/documents/ref/TRD_2012_02_E.pdf.

Martell S., and Stewart I. 2013. Towards defining good practices for modeling time-varying selectivity. Fisheries Research, 158: 84–95.

Maunder M. N., and Punt A. E. 2013. A review of integrated analysis in fisheries stock assessment. Fisheries Research, 142: 61–74.

Methot R. D., and Wetzel C. R. 2013. Stock synthesis: providing a biological and statistical framework for fishery management forecasts across a data-poor to data-rich continuum. Fisheries Research, 142: 86–99.

Mohn R. 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES Journal of Marine Science, 56: 473–488.

Myers R. A., Bowen K. G., and Barrowman N. J. 1999. Maximum reproductive rate of fish at low population sizes. Canadian Journal of Fisheries and Aquatic Sciences, 56: 2404–2419.

NOAA. 2009. Report of the retrospective working group. Northeast Fisheries Science Center Reference Documents, 09-01. National Oceanic and Atmospheric Administration, Woods Hole, MA. http://www.nefsc.noaa.gov/nefsc/publications/crd/crd0901/crd0901.pdf.

Ono K., Licandeo R., Muradian M. L., Cunningham C. J., Anderson S. C., Hurtado-Ferro F., Johnson K. F., *et al.* 2014. The importance of length and age composition data in statistical age-structured models for marine species. ICES Journal of Marine Science, doi:10.1093/icesjms/fsu007

Parma A. M. 1993. Retrospective catch-at-age analysis of Pacific halibut: implications on assessment of harvesting policies. *In* Proceedings of the International Symposium on Management Strategies for Exploited Fish Populations, pp. 247–265. Alaska Sea Grant College Program.

R Core Team. 2014. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Sinclair A., Gascon D., Rivard D., and Gavaris S. 1991. Consistency of some northwest Atlantic groundfish stock assessments. NAFO Scientific Council Studies, 16: 59–77.

Smith A. D. M., Sainsbury K. J., and Stevens R. A. 1999. Implementing effective fisheries-management systems—management strategy evaluation and the Australian partnership approach. ICES Journal of Marine Science, 56: 967–979.

TRAC (Transboundary Resources Assessment Committee). 2012. Stock assessment of Georges Bank Yellowtail Flounder for 2012. TRAC, Status Report 2012/02. http://www2.mar.dfo-mpo.gc.ca/science/trac/rd.html.

Valero J. L. 2012. Harvest policy considerations on retrospective bias and biomass projections. IPHC Report of Assessment and Research Activities, 2011: 311–330.

Handling editor: Mark Maunder