NRC Research Press

ARTICLE

Using management strategy evaluation to establish indicators of changing fisheries[1]

T.R. Carruthers and A.R. Hordyk

Abstract: A new indicator is described that uses multivariate posterior predictive data arising from management strategy evaluation (MSE) to detect operating model misspecification (exceptional circumstances) due to changing system dynamics. The statistical power of the indicator was calculated for five case studies for which fishery stock assessments have estimated changes in recruitment, natural mortality rate, growth, fishing efficiency, and size selectivity. The importance of the component data types that inform the indicator was also calculated. The indicator was tested for multiple types of management procedures (e.g., catch limits by stock assessment, size limits, spatial closures) given varying qualities of data. The statistical power of the indicator could be high even over short time periods and depended on the type of system change and quality of data. Statistical power depended strongly on the type of management approach, suggesting that indicators should be established that rigorously account for feedbacks between proposed management and observed data. MSE processes should use alternative operating models to evaluate protocols for exceptional circumstances to ensure they are of acceptable statistical power.

Résumé : Un nouvel indicateur est décrit qui fait appel à des données de distributions multivariées prédites a posteriori issues de l'évaluation de stratégies de gestion (ESG) pour déceler des erreurs de spécification (des circonstances exceptionnelles) dues à des changements de la dynamique du système. La puissance statistique de l'indicateur a été calculée pour cinq études de cas pour lesquelles des évaluations de stocks de ressources halieutiques ont estimé les variations du recrutement, du taux de mortalité naturelle, de la croissance, de l'efficacité de la pêche et de la sélectivité selon la taille. L'importance des types de données utilisées pour déterminer l'indicateur a également été calculée. L'indicateur a été testé pour différents types de procédures de gestion (p.ex. limites de prises découlant d'évaluation des stocks, limites de taille, fermetures spatiales) selon la qualité des données. La puissance statistique de l'indicateur peut être grande même sur de courtes périodes de temps et dépend du type de changement du système et de la qualité des données. La puissance statistique dépend fortement du type d'approche de gestion, ce qui porte à croire que des indicateurs devraient être établis qui tiennent rigoureusement compte des rétroactions entre la gestion proposée et les données observées. Les processus d'ESG devraient faire appel à différents modèles d'exploitation pour évaluer les protocoles visant les circonstances exceptionnelles pour vérifier que leur puissance statistique est acceptable. [Traduit par la Rédaction]

Introduction

Management strategy evaluation (MSE) is increasingly adopted in fisheries to identify management procedures (MPs, aka harvest strategies: algorithms that calculate management recommendations from data) that are robust to uncertainty in system dynamics and observation processes (Butterworth and Punt 1999; Cochrane et al. 1998). MSE relies on operating models that simulate the fishery system, including fishing dynamics, population dynamics, observation processes, and the implementation of management recommendations. By using a simulated system, the expected performance of MPs can be quantified over a wide range of uncertainties while accounting for feedbacks among the advice arising from MPs, the system dynamics, and the data. The MSE approach allows managers to quantify performance trade-offs, identify robust MPs that can achieve desirable management outcomes, and establish the value of various data-collection programs.

Current best practice in MSE is to derive operating models empirically by fitting them to data (Punt et al. 2016). Although decisions over operating model structure and metrics of management performance often have subjective components, once these are established the selection of a suitable MP arises objectively from MSE simulations. A principal benefit of MSE is that MPs are generally easier to understand for a wider range of stakeholders, and the rationale for their selection is transparent and defensible (Butterworth 2008). By establishing an MP, stakeholders of the resource agree to a common set of rules for future management.

A principal motivation behind the MP approach was originally to lessen the need for frequent use of more complex and comprehensive stock assessment processes (and associated "tinkering"; Butterworth 2008). It follows that in most settings, MPs are adopted for an agreed period of time after which a formal review of the MP is scheduled (e.g., the 5-year interval for reviews of MP implementation by the International Whaling Commission 1999). There is, however, a need to establish protocols for detecting situations where the observed system dynamics are not consistent with the range of simulations specified in the operating model, over which the MP was demonstrated to be robust (Butterworth 2008). Referred to as "exceptional circumstances", these protocols

Received 4 June 2018. Accepted 1 August 2018.

T.R. Carruthers and A.R. Hordyk. Institute for the Oceans and Fisheries, Aquatic Ecosystems Research Laboratory (AERL), 2202 Main Mall, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Corresponding author: Thomas R. Carruthers (email: t.carruthers@oceans.ubc.ca).

[1]This article is being published as part of the special issue "Under pressure: addressing fisheries challenges with Management Strategy Evaluation" arising from two related theme sessions sponsored by the American Institute of Fishery Research Biologists at the 147th Annual Meeting of the American Fisheries Society, Tampa, Florida, USA, August 2017.

are often described in detail but the statistical justification for them has been generally lacking. For example, in the case of Greenland halibut (*Reinhardtius hippoglossoides*), the Northwest Atlantic Fisheries Organization (NAFO 2010) compares observed catches and survey indices with operating model predictions of these data (posterior predicted data). When observations fall out of the 90% probability interval of posterior predicted data, these are considered indicative of exceptional circumstances, implying a rate of false positives (incorrectly rejecting the operating model) of 10%. Similarly, Fisheries and Oceans Canada (DFO 2011) identifies various data that may be monitored to detect exceptional circumstances in the MSE for pollock (*Pollachius virens*) in NAFO area 4Xopqrs5, including survey biomass indices, catches, and catch-at-age composition data. DFO (2011) also identifies the 90% probability interval of simulated survey data as a possible indicator and also includes an absolute level for the posterior predicted indices below which observed data would indicate exceptional circumstances. Exceptional circumstances provisions were triggered for Greenland halibut (NAFO 2018) in 2014 when one of the survey index observations fell below the fifth percentile of posterior predicted distributions. Suppose, however, this had not been triggered; what assurances are there that the proposed indicator would be effective in detecting departures from the assumed operating model? In both the halibut and pollock examples, it is not known how frequently such indicators can be expected to correctly detect a departure from the simulated operating model conditions (the statistical power of the indicator). In the wider context of ecological monitoring, it has been argued that power analysis is essential in determining suitable indicators (Peterman 1990; Murtaugh 1996).

There are two possible responses to the triggering of exceptional circumstances in an MSE process. One is to bring forward a formal review of the MP, potentially including a new MSE analysis from which a new MP is selected. The other involves an ad hoc adjustment to the recommendation of the management procedures (Butterworth 2008). The advantage of the second approach is that it can be codified in the MP and therefore tested by MSE. For example, de Moor et al. (2011) used MSE to test MPs for the South African sardine (*Sardinops sagax*) and anchovy (*Engraulis encrasicolus*) fishery that explicitly included exceptional circumstances in their provision of management advice. In this case, the maximum reduction in catch recommendations of the MP was overridden, and larger reductions were permitted in cases where the survey biomass estimates dropped below a specified threshold. The explicit inclusion of exceptional circumstances provisions within the MPs has also been proposed for Pacific tropical tuna fisheries by Scott et al. (2016).

MSE is an expanding field, and most existing MSE processes have been implemented relatively recently (the last 15 years). Perhaps for these reasons there has been a focus on establishing operating models and adopting MPs. For consistency, protocols for rejecting an operating model in use, and potentially the MP established by it, should have similar transparency, accountability, and objectivity as those applied in operating model specification and MP adoption. Identifying statistically rigorous indicators of operating model misspecification may be more important in MSE processes for data-limited stocks where the statistical power of indicators can be expected to be lower due to greater uncertainty in operating model dynamics and the availability of less numerous, precise, and accurate data.

We distinguish between two types of operating model misspecification. The first is an operating model that misspecifies both historical and future system dynamics (i.e., the operating model is not a representative description of the current and historical fishing and stock dynamics). The second type of model misspecification occurs due to unforeseeable changes in future system dynamics, such as recruitment strength, natural mortality rate, somatic growth, fishing selectivity, or fishing efficiency, that were not accounted for in the operating models. In this paper, we focus on the latter type and assume that operating models were correctly specified but changes occurred in future system dynamics after an MP has been adopted and is in use (although both types of misspecification can be tested with the approach described here).

We extend existing simplistic approaches for operating model rejection and use multivariate posterior predicted data generated by an operating model to identify indicators that have sufficient power to detect when operating models no longer provide a suitable representation of the fishery system. To ensure that the type and magnitude of system changes are plausible, we derived these from fishery stock assessments that have reported such shifts. These case studies are used to both evaluate the statistical power of the indicator approach and identify which types of data are most informative of the various changes in system dynamics for various classes of MP. The objective of this paper is to evaluate a general approach to establishing indicators of operating model misspecification due to temporal shifts in parameters rather than to rigorously derive indicators for the specific case studies and MPs of this paper. We focus on the first step of identifying suitable indicators of operating model misspecification that may be later codified in MPs as exceptional circumstances and tested with MSE (e.g., de Moor et al. 2011).

Methods

Measuring the similarity between posterior predictive data and real observations

MSE involves closed-loop simulation, where fishery data are simulated based on the operating model, an MP is applied to the simulated data to produce a management recommendation, and the system dynamics in the next time step (usually annual) are updated based on the implemented management recommendation. This process is repeated for each MP evaluated in the MSE. It follows that the simulated fishery data in the future projection years arise from the application of a particular MP (posterior predictive data).

The central question is whether the posterior data predicted by operating models at the time of MP adoption can be used to test the suitability of the operating model in future years as an MP is used and real data from the fishery system are observed. The general problem is one of quantifying the similarity of observed data in comparison with many simulations of MSE posterior predictive data.

Preliminary investigation of possible indicators (data types) for rejecting operating models identified a number of design considerations. Any such indicator cannot rely on a particular parametric model for the multivariate distribution of posterior data. Various time series of data such as mean length of catches, catch per unit of effort (CPUE), and magnitude of annual catches are often distributed irregularly and are difficult to approximate by parametric models. Another constraint is that the calculation of similarity between observations and posterior predictive data must be efficient for high dimensional data. Even if only three types of data are considered, repeat observations of these data in future years increase the total number of observations linearly.

A standard measure of the discrepancy between a point (e.g., a datum in multivariate space) and a distribution (the posterior predicted data of an operating model) is the Mahalanobis (1936) distance D and is defined as follows:

$$D = \sqrt{(\vec{d} - \vec{\mu})^T \mathbf{C}^{-1} (\vec{d} - \vec{\mu})} \quad (1)$$

where $\vec{d}$ is a vector of data ($\vec{d} = (d_1, d_2, d_3, \ldots, d_N)$), $\vec{\mu}$ is a vector of mean values for those data ($\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)$), and $\mathbf{C}$ is the covariance matrix of the data. For example, d_1 may be mean catches after 6 years, d_2 may be mean length of catches after

Fig. 1. The process of accepting or rejecting operating models for observed data (e.g., data types d1 and d2) based on the Null operating model (steps 1–4) and the calculation of statistical power (1 – beta) using an Alternative operating model (steps 5 and 6). [Colour online.]

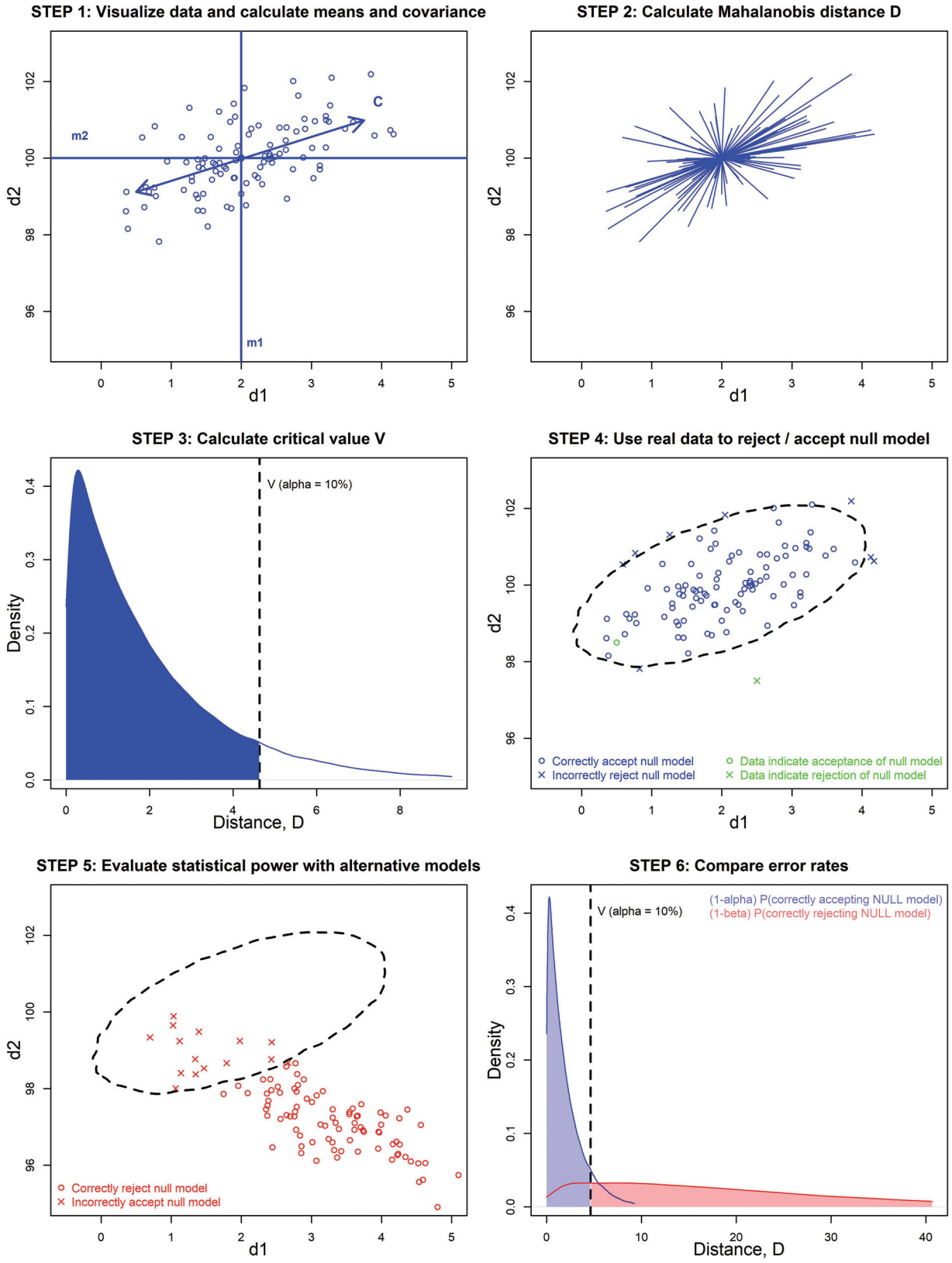


Table 1. Data types used in the indicator.

Data type	Code	Details
Mean catches	C_m	Mean of log catches
Interannual variability in catches	C_v	$\frac{1}{n}\sum_{t=y}^{y+n-1}\frac{\lvert C_{t+1}-C_t\rvert}{C_t}$
Slope in catches	C_s	Maximum likelihood estimate of slope of log catches
Slope in mean length in catches	ML_s	Maximum likelihood estimate of slope of log mean length
Mean length in catches	ML_m	Mean of log length
Mean abundance index	I_m	Mean of log index
Slope in abundance index	I_s	Maximum likelihood estimate of slope of log abundance index

6 years, d_3 could be biomass survey index between years 7 and 12, and so on. In this example, μ_1 is the mean of the posterior catch from the MSE after the first 6 future projection years, and so on for the other elements in $\vec{\mu}$.

Mahalanobis distance is similar to standard Euclidean distance but accounts for parameter cross-correlation via the covariance matrix **C**. This is important in this context, as repeat observations in time (e.g., annual catches) and among data (e.g., catch and CPUE) can be expected to be correlated. When the covariance matrix is the identity matrix, Mahalanobis distance is equal to Euclidean distance. Mahalanobis distance is used widely in related multivariate analyses, including outlier detection (Gnanadesikan and Kettenring 1972; Leys et al. 2018), principal component analysis (Brereton 2015), and calculation of multivariate probability density (Rose and Smith 1996).

The indicator approach tested here has the following steps (Fig. 1):

(1) For S simulations, N multiple posterior data types (d_1, d_2, …, d_N) are calculated, and the means (μ_1, μ_2, …, μ_N) and covariance **C** of these data are calculated.
(2) For each simulation, Mahalanobis distance D is calculated (eq. 1), creating a set of possible distances $\theta = \{D_1, D_2, D_3, \ldots, D_S\}$.
(3) Based on a specified rate of incorrectly rejecting the Null operating model alpha, a percentile (1 – alpha) of θ is calculated that is treated as a critical value V.
(4) The distance $\hat{D}$ is then calculated for real data; where $\hat{D} < V$ the operating model is accepted, and where $\hat{D} > V$ the operating model is rejected.

A numerical approach to evaluate statistical power of the indicator

To establish the statistical power of the indicator, pairs of operating models were specified. The Null model includes constant system dynamics; the Alternative model includes a plausible shift in future system dynamics estimated in a real fishery stock assessment. For each pair of operating models, the indicator is established as in steps 1–3 above using the Null model. However, distances are also calculated for the simulated data of the Alternative model (Step 5, Fig. 1). Rather than a single distance for a single data observation, this produces a set of distances for the Alternative model data from the Null model data; $\hat{\vartheta} = \{\hat{D}_1, \hat{D}_2, \hat{D}_3, \ldots, \hat{D}_S\}$.

Assuming that the Alternative model is true, the rate of failing to reject the Null operating model (beta) is represented by the fraction of $\hat{\vartheta}$ distances less than V (Step 6, Fig. 1). Assuming the Null hypothesis is true and no system changes took place, the rate of falsely rejecting the Null operating model is determined by alpha. As increasingly smaller changes in system dynamics are simulated by the Alternative model, the Null and Alternative models become similar; $\hat{\vartheta}$ and ϑ therefore become similar and beta tends to 1 – alpha. In these analyses we follow the example of Greenland halibut (NAFO 2010) and assumed an alpha of 10%.

Posterior data types

There are numerous possible data and derived quantities that may be used to detect operating model misspecification. In theory these could include annual observations of catches, survey indices, CPUE indices, size composition, etc. However, if the objective is to detect departure in system dynamics, it may be more appropriate to monitor derived quantities of higher signal–noise ratio, such as mean trend in survey indices or average catches over several years.

We identify seven types of derived data monitored over future intervals (Table 1): mean catches (C_m), interannual variability in catches (C_v), slope in catches (C_s), slope in mean length in catches (ML_s), overall mean length in catches over several years (ML_m), the mean abundance index (I_m), and the slope in the abundance index (I_s). Each type of data is calculated over 6-year intervals of the MSE projection, years 1–6, 7–12, 13–18, etc. In each future time step, the latest seven data types are added to the indicator, such that the dimensionality of the indicator increases by seven after each future time increment of 6 years. Time blocks of 6 years were chosen arbitrarily to demonstrate the indicator concept for the purposes of this paper. This length of time block was sufficient to lessen the noise in data of high interannual error and obtain estimates of slope in data while offering an assessment of the operating model over the near term (within the first 6 years of MP adoption).

It is possible that data observations may be linearly predicted by a combination of other data types. For example, an MP that specifies a fixed level of effort may lead to CPUE and catches that are highly correlated. In such cases the covariance matrix **C** may be singular and noninvertible, preventing the calculation of D in eq. 1. To overcome this, we used singular value decomposition (Golub and Van Loan 1989) to calculate the "pseudo inverse" correlation matrix $\mathbf{C}^{-1}$.

A principal limitation of the approach described here is that it assumes data are approximately unimodal. If, for example, posterior predictive data are strongly and symmetrically bimodal, the data from an Alternative model could be closer to the mean of the Null data (smaller distances) and yet uncharacteristic of the data from the Null model (in this example there is low posterior density between two modes). In such cases the indicator could have very low statistical power. To prevent this, a check was included; if any data in any time step were not approximately unimodal, they were rejected and not used in the calculation of distances. Hartigan's (1985) test of unimodality was used; posterior data obtaining a Hartigan's statistic higher than 0.05 were not used by the indicator (however, this was not triggered in any of the simulations).

Observation models

Since the statistical power of the proposed indicator approach is likely to be strongly dependent on the quality of the simulated posterior data, it was important to include alternative observation models that generate data of varying frequency, bias, and precision. In this case the data used by the MPs and the indicator include annual catches, survey indices, and length composition data (to calculate mean lengths).

Two observation models ("good" and "bad") were applied that were intended to represent observation processes of a data-rich

Table 2. Observation error models used to generate simulated posterior predicted data (range shown in parentheses).

Observation model quantity	Good (data-rich)	Bad (data-poor)
Lognormal standard deviation of annual catches	(0.1, 0.2)	(0.2, 0.6)
Lognormal standard deviation of mean bias in annual catches	0.05 (approximately 95% of simulations between −10% under-reporting and +10% over-reporting)	0.3 (approximately 95% of simulations between −60% under-reporting and +80% over-reporting)
Lognormal standard deviation of annual survey index observations	(0.1, 0.25)	(0.2, 0.6)
Hyperstability–hyperdepletion in the annual survey index	(2/3, 3/2) (a true 50% decline in vulnerable biomass is represented as a 37% to 75% decline by the index)	(1/3, 3) (a true 50% decline in the vulnerable bias is represented as a 20% to 85% decline by the index)
No. of independent annual length observations	(50, 100)	(10, 20)

Table 3. Management procedures applied in indicator testing.

Management procedure	Code	Type	Rationale–role	Reference
Delay-difference assessment	DD	TAC	Represents a simple data-rich stock assessment using time series data recommending a TAC at F_{MSY} fishing rates (e.g. TAC = current vulnerable biomass × F_{MSY})	Carruthers and Hordyk 2018
Delay-difference assessment with effort control	DDe	TAE	As with DD, but recommends F_{MSY} effort levels	
Fishing selectivity matches maturity	matlenlim	Minimum size limit	Static management with compensatory attributes	—
Current effort	curE	TAE	Status quo fishing	—
Marine reserve (10% habitat with effort reallocation)	MRreal	Spatial closure	—	—
Mean-catch depletion	MCD	TAC	A robust data-moderate MP that generally outperforms depletion-corrected average catch and depletion-based stock reduction analysis (Dick and MacCall 2011)	Harford and Carruthers 2017

Note: TAC, total allowable catch; TAE, tota allowable effort.

and a data-limited fishery, respectively. The parameters of these observations models are provided in Table 2. The purpose of these alternative models is to demonstrate the potential impact on statistical power — these observation models are not intended to provide a credible recreation of observation processes for the various case studies.

Case studies of changing system dynamics

Five case studies with documented stock assessments were identified where historical shifts in system dynamics have been documented (Fig. 2; see Appendix A for details). These case studies include a roughly twofold increase in recruitment strength in approximately 1986 for East Atlantic bluefin tuna (*Thunnus thynnus*) (Anonymous 2018; Carruthers 2018*a*), a threefold increase in natural mortality rate in 1996 for Atlantic cod (*Gadus morhua*) in NAFO area 4x5Y (Clark and Emberley 2010; Carruthers 2018*b*), an average 9% annual increase in fishing efficiency in Faroe Islands haddock (*Melanogrammus aeglefinus*) (Eigaard et al. 2011, Hordyk 2018*a*), a fourfold increase in growth rate and a decrease in asymptotic length for Pacific hake (*Merluccius productus*) in 1989 (Stewart et al. 2011; Hordyk 2018*b*), and a shift in gear selectivity towards older, larger individuals in Gulf of Mexico vermillion snapper (*Rhomboplites aurorubens*) (SEDAR 2016; Carruthers 2018*c*). In each case, the Null operating model is specified using the best available information for the stock at that time and therefore does not include the unexpected changes in the future dynamics. The Alternative operating model assumes that the estimated changes occurred and were perpetuated into future years. Since some of the alternative operating models include relatively pronounced shifts, Intermediate models were also parameterized to evaluate the indicator's ability to detect lesser changes. The attributes of these various operating models are presented in Fig. 2.

Example management procedures

Multiple management procedures were included to test the indicator approach under management systems of varying responsiveness, data dependencies, and aggressiveness (Table 3; see Appendix B for details). These MPs include total allowable catch (TAC) recommendations by simple data-rich delay-difference assessments (DD, DDe), a size limit (matlenlim), total allowable effort (TAE) set to status quo current fishing effort (curE), a 10% marine reserve (MRreal), and a robust data-moderate MP (mean-catch depletion, MCD). It was necessary to test a diversity of MPs, since posterior data may have varying information to detect model misspecification depending on the status of the resource and the variability or magnitude of exploitation.

Rating MP performance

Simply knowing that the indicator was insufficiently powerful to detect a real shift in system dynamics does not address whether this failure was consequential. To evaluate this, a "best" MP was selected for each operating model based on the ability to keep population biomass close to the level at maximum sustainable yield (B_{MSY}, commensurate with sustainable and substantial yields). Over all 200 simulations and 30 projected years, the probability of biomass (PB) staying between 50% and 150% B_{MSY} was calculated. For each operating model, the highest performing MP was selected to evaluate whether Alternative and Intermediate operating models led to the selection of a differing MP of contrasting performance.

Quantifying the contribution of data to the indicator

Garthwaite and Koch (2016) derive an approach for calculating the relative contribution of the various data types to the Mahalanobis distance used by the indicator. Recall that for a vector of simulated data from an Alternative operating model $\vec{d}$ that differs

Fig. 2. Case studies used to test statistical power of the indicator approach. Blue lines represent the Null operating model with no changes. Solid red lines represent the Alternative operating model with the full extent of change estimated by the various stock assessments. The dashed red line represents the Intermediate operating model that exhibits half the extent of system change. The leftmost dashed vertical line represents the end of the historical period after which changes were estimated by the stock assessments. The second vertical line marks the end of estimation period after which the system dynamics of the Null, Alternative, and Intermediate operating models are assumed to continue over further projection years. Panel *a* also includes the maximum likelihood estimate of the recruitment strength from the assessment model (a solid black line). The selectivity changes for vermilion snapper (panel *e*) are expressed as the mean age in the catch that would arise from the time-varying selectivity if it were applied to an unfished population. [Colour online.]

from the Null operating model means $\vec{\mu}$, the Mahalanobis distance is calculated by eq. 1 using the covariance matrix **C**. The contribution Ω of the *j*th data type $\vec{d}_j$ to D can be calculated as

$$\Omega_j = \left[\mathbf{YCY}^{-1/2}\mathbf{Y}(\vec{d}_j - \vec{\mu})\right]^2 \quad (2)$$

where **Y** is a diagonal matrix such that **YCY** has diagonal entries equal to 1 and is found by the "corr-max" transformation (see Garthwaite and Koch 2016 for derivation of eq. 2 and Garthwaite and Koch 2018 for a worked example; functions for this calculation are included in the R package MSEtool Carruthers et al. 2018). For a given posterior sample of data *i*, the contribution of data type *j* was standardized as a percentage of all contributions: $\phi_{i,j} = 100\Omega_{i,j}/\sum_j \Omega_{i,j}$. These contributions of the various data types *j* can then be summarized over all posterior samples according to means and percentiles of the relative contribution. These contributions can be used to evaluate which types of data were most critical in determining overall Mahalanobis distance for each type of change in system dynamics represented by the various case studies.

A general concern of the indicator approach outlined above is that it might include data that are not indicative of differences between the Null and Alternative operating models, thereby reducing the power of the indicator due to additional noise. The calculation of ϕ allowed for the derivation of a second indicator that ignored data types that did not substantially contribute to the distance between Null and Intermediate or Alternative operating models. This second indicator only used those data types contributing to the top 95% of the total Mahalanobis distance whenever it was calculated.

MSE framework

All MSE calculations were conducted using the R packages DLMtool (Carruthers and Hordyk 2018) and MSEtool (Carruthers et al. 2018), the latter containing open-source functions for the indicator approach of this paper. The operating models applied in this research are fully documented and available online at DLMtool (2018). For each MSE, 200 simulations were conducted to characterize the multivariate posterior predicted data for the various operating models.

Results

The statistical power of the indicator varied among case studies, MPs, operating models, and update years (Fig. 3). In general, the statistical power was highest for the natural mortality rate shifts of cod, followed by the growth changes of hake, the recruitment increases of bluefin tuna, with substantially lower (and more variable among MPs) power for the catchability increases of haddock and the selectivity change of vermilion snapper (Fig. 3). The impact of data quality on statistical power was greater than the degree of shift in system dynamics represented by the Intermediate or Alternative operating models (Figs. 3 and 4). In the case of vermilion snapper, even after 30 years the indicator did not provide statistical power greater than the threshold level of 80% regardless of the MP applied and the quality of the data (Fig. 3).

In most cases, additional time blocks of data increased the statistical power of the indicator (Figs. 3 and 5). The absolute statistical power and the increase in power as time blocks of data were added varied substantially among the MPs for all case studies. For example, in the haddock case study, management by size limit (matlenlim) led to strongly differing posterior data among Null and Alternative operating models and therefore higher statistical power than the DD MPs (Fig. 3 panels *c* and *m*). MPs that produced data with relatively high power to detect shifts in one case study could provide very little power to detect the shifts of another case study (e.g., matlenlim applied to hake and vermilion snapper; Fig. 3 panels *n* and *o*).

Robust MPs that perform better under alternative operating models may be expected to provide less contrast in fishery data. This is confirmed in Fig. 3 where MPs obtaining higher probability of biomass between 50% and 150% of B_{MSY} over the 30-year projection were generally those with the lowest statistical power (e.g., the DD MP for haddock).

Notable exceptions to the findings above are the poor performance of the MCD MP for haddock that also had weak statistical power. In this particular case, the cause was chronic underfishing (91% of simulations were higher than $1.5B_{MSY}$), which also suggests, perhaps not surprisingly, that the power of the indicator may be driven more strongly by fishery collapses.

At first glance the results of the hake operating models appear counterintuitive, with Intermediate operating models achieving higher statistical power than the stronger Alternative operating models. In the Alternative operating models, the shift to lower asymptotic length is matched by an upward shift in growth rate (the von Bertalanffy *K* parameter). The Intermediate operating model split the difference in asymptotic length but kept the full increase in *K* leading to a more dramatic shift in size composition than simulated with the Alternative operating model. This example demonstrates that the impact of system changes in one aspect is often not separable from other related quantities.

The selection of the "best" MP changed among the Null, Intermediate, and Alternative operating models (Table 4), with the exception of haddock with good quality data and hake with bad quality data. In the case of haddock, the fishing efficiency changes had no impact on the performance of TAC control using the DD MP. The finding that MP selection is changeable is important, as it underlines that the relative lack of statistical power in case studies such as vermilion snapper may mask the need for formal MP evaluation.

For the case of the delay-difference MP providing TAC recommendations (DD) for cod, it is clear that certain data types contribute much more strongly to the discrepancy between posterior predicted data of the operating models (Fig. 4). In this case, the most important contributor is the slope in the survey index (*I*_s). This is reflected in Table 5, where the mean contribution of *I*_s is as high as 63% of the indicator for the first 6 years. Interestingly, these data become unimportant where the data for the second time block (years 7 and 12) now dominate the contribution to the indicator and contribution is spread more evenly among the various data types.

The contribution of the various data varied strongly among the case studies and operating models within case studies. For example, over the first 6 years, mean length data were the most contributory to the indicator for hake with bad quality data (55% and 44% contribution for the Intermediate and Alternative operating

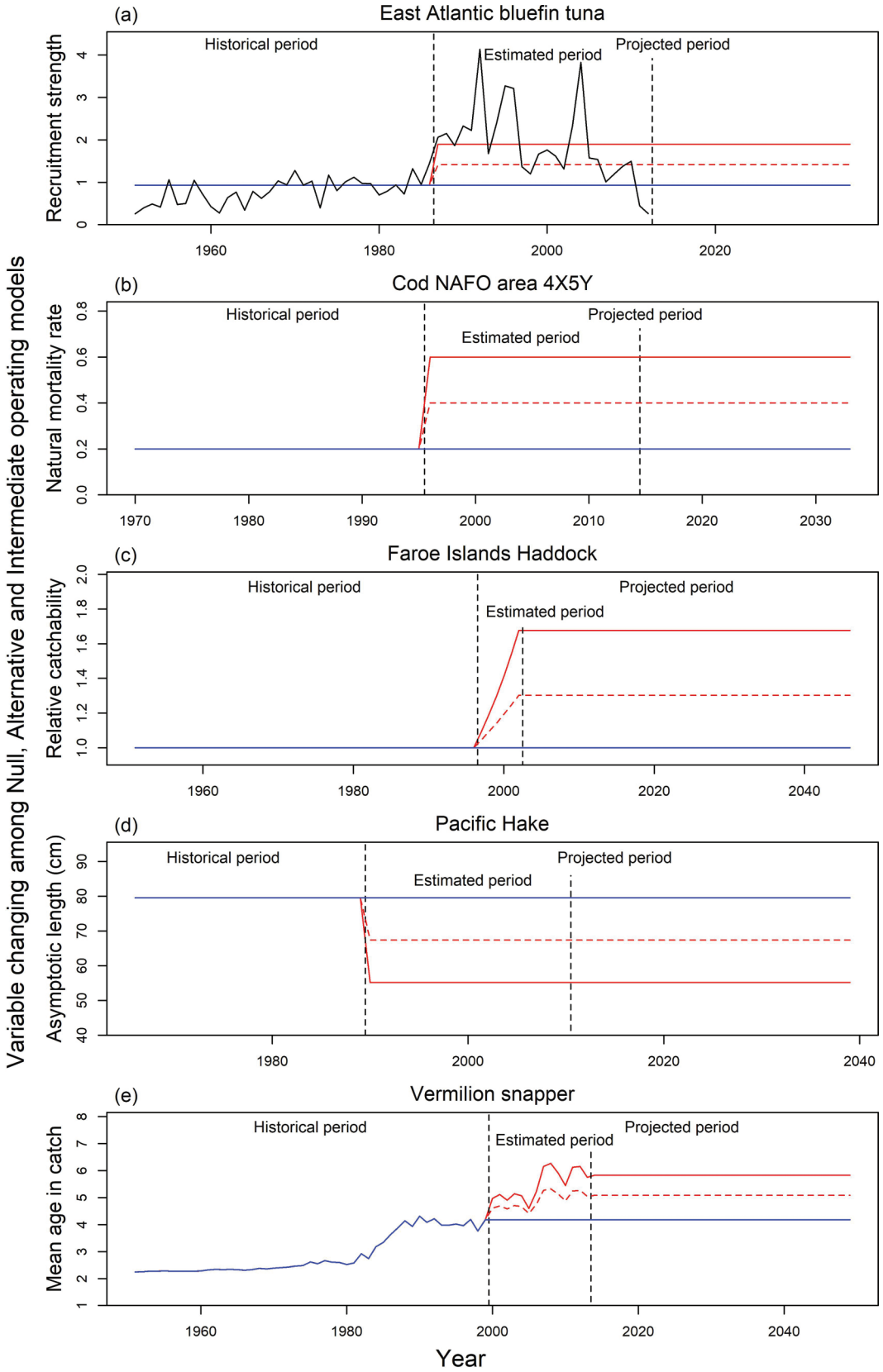
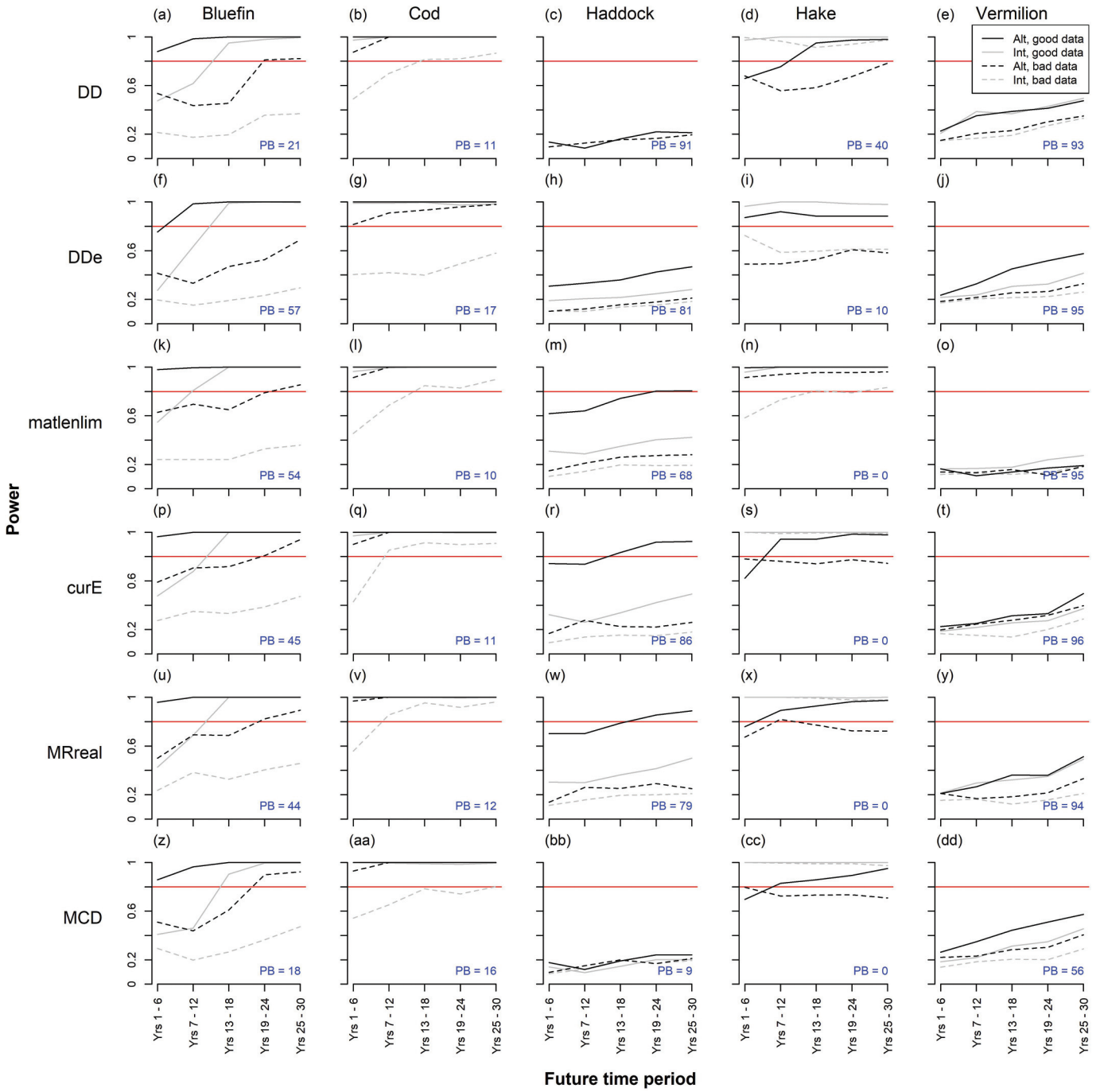


Fig. 3. The statistical power of the indicator for all case studies and operating models over five time blocks of 6 years. The red horizontal line represents a reference power level of 80%. The blue number in each panel is the MP performance metric PB that is the probability of biomass between 50% and 150% of B_{MSY} throughout the 30-year projection in the case of the Alternative operating model with good data. [Colour online.]



models, respectively). However, this was not the case for the good observation model, where variability in catches was a larger contributor (*C*_v; Table 5).

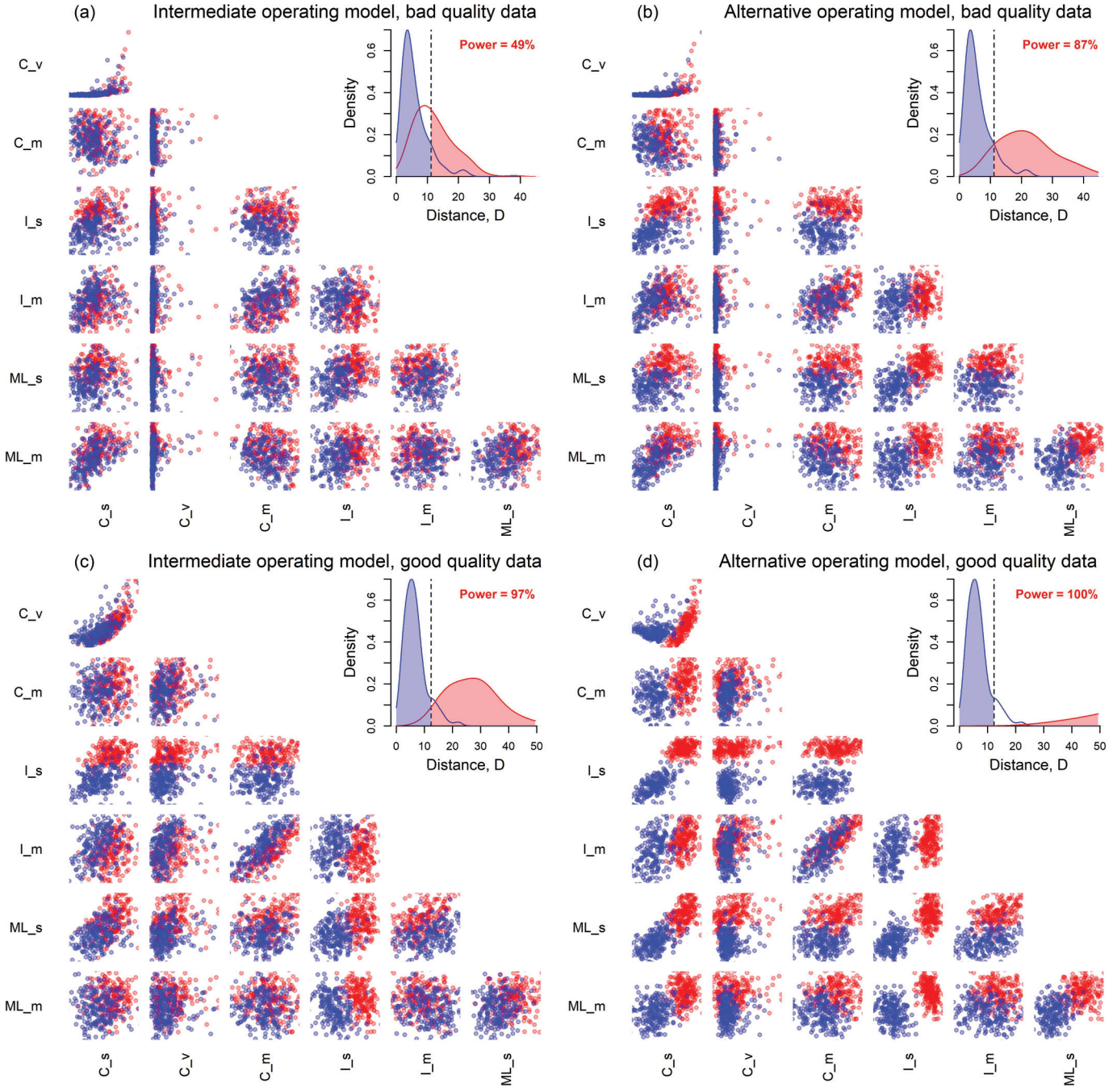
The contribution of data to the indicator was also dependent on the type of management procedure being applied. The delay-difference assessment providing total allowable effort (TAE) recommendations (DDe) (Table 6) led to an indicator that was driven by different data types than the TAC-based DD MP (Table 5). In the example of hake, for the first time block of 6 years, the slope in annual catches (*C*_s) was not important for the indicator when the DD MP was applied (4%–9% contribution; Table 5) but important for the DDe MP (15%–28% contribution; Table 6).

A version of the indicator in which differences among the posterior predictive data were limited to only the most contributory data types (the top 95% of contributions) did not provide substantially higher statistical power (Fig. 6) than an indicator including all data types (Fig. 3).

Discussion

The objective of this paper was to evaluate an approach to developing indicators of fishery system change that could be used to reject operating models once MPs are in use (exceptional circumstances). A new indicator using multivariate posterior predictive

Fig. 4. Posterior predictive data from the cod operating models, calculated over the first 6 MSE projection years for the delay-difference MP (DD). The blue points represent the posterior predicted data of the Null model. The red points represent the posterior predicted data of the Intermediate (half the change in dynamics of the Alternative operating model) (*a* and *c*) and Alternative (*b* and *d*) operating models. Posterior data were generated for the bad-quality (*a* and *b*) and good-quality (*c* and *d*) observation models. The cross-correlation of seven data types are presented that include the variability (_v), median (_m), and slope (_s) in annual catches (*C*), a fishery-independent abundance index (*I*), and mean length (ML) calculated over the first 6 projected MSE years. The statistical powers calculated in these four plots are represented as the start of the four plotted lines in Fig. 3*b*. [Colour online.]



data was proposed that is a multivariate extension of current protocols for exceptional circumstances (e.g., DFO 2011; NAFO 2010). As proof of concept, the indicator was tested by statistical power analysis for five types of system change (case studies), two data qualities, and two magnitudes of system change. The results suggest that the indicator can successfully detect changes in fishery system dynamics even over the first 6 years of MP adoption. However, there are few general rules, and the statistical power of the indicator depended strongly on the quality of observed data, the management procedure in use, the current fishery dynamics, the type of potential change in fishery dynamics, and the time horizon over which the data will be used. A key recommendation arising from these findings is that any such indicator should be tested on a case-specific basis using candidate MPs and

Fig. 5. Power analysis of the indicator approach given the Alternative operating model for East Atlantic bluefin tuna for the six MPs over five future time blocks of 6 years. Blue and red distributions represent the Mahalanobis distances of the Null and Alternative operating models, respectively. [Colour online.]

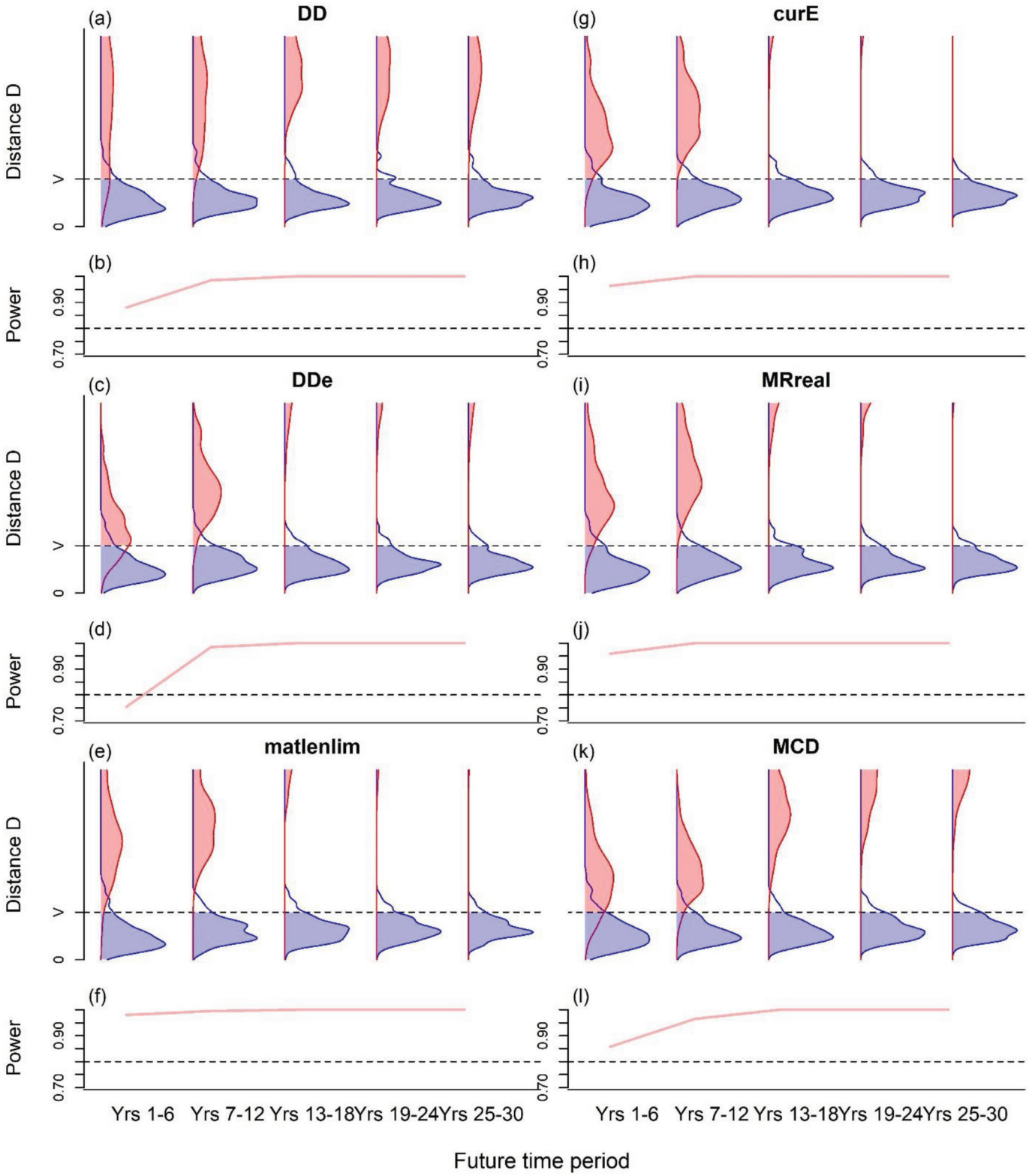


Table 4. The selection of "best" MPs by operating model.

Observation model	Operating model	Bluefin tuna		Cod		Haddock		Hake		Vermilion snapper	
Good	Null	matlenlim	100	DD	52	DD	91	DDe	27	matlenlim	97
	Intermediate	matlenlim	88	MCD	33	DD	91	DD	28	curE	98
	Alternative	DDe	57	DDe	17	DD	91	DD	40	curE	96
Bad	Null	matlenlim	100	DD	45	DD	61	DD	15	matlenlim	97
	Intermediate	matlenlim	88	MCD	33	curE	61	DD	20	curE	98
	Alternative	DDe	55	MCD	18	curE	86	DD	27	curE	96

Note: The MP selected obtained the highest probability of keeping biomass between 50% and 150% of biomass at maximum sustainable yield over the 30-year projection period (i.e., a high probability of achieving a productive stock size). The numbers are the percent probabilities.

Table 5. The mean percent contribution (ϕ) of the various data to the indicator for the delay-difference model MP controlling total allowable catches (DD) after 6 and 12 update years.

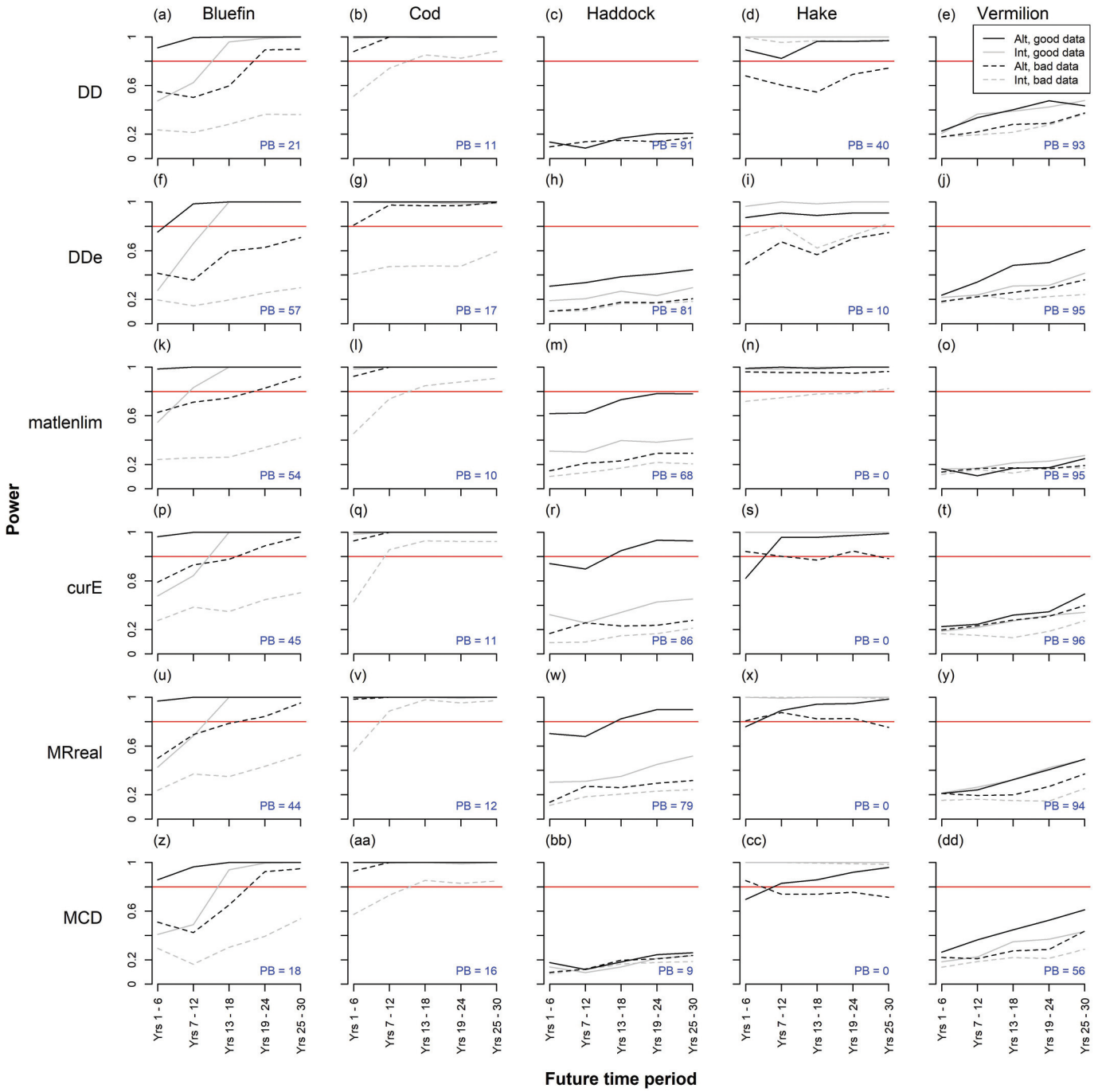
			Data type: first time block (years 1–6)							Data type: first two time blocks (years 1–6, 7–12)												
			Calculated over projection years 1–6							Calculated over projection years 1–6							Calculated over projection years 7–12					
Case study	Data quality	Operating model	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m
Bluefin tuna	Bad	Intermediate	13	1	14	11	20	20	21	7	1	8	5	10	11	11	6	5	9	7	7	6	8
		Alternative	9	1	14	10	20	23	23	5	1	6	5	12	14	15	5	4	9	5	7	4	8
	Good	Intermediate	9	8	16	16	29	10	12	4	4	5	6	21	5	4	4	4	15	5	12	4	5
		Alternative	6	4	19	11	42	10	7	2	2	5	4	33	4	4	1	3	21	4	12	3	2
Cod	Bad	Intermediate	8	1	16	25	22	14	14	3	1	5	7	12	6	6	10	3	13	8	9	6	11
		Alternative	5	1	15	39	12	15	13	1	1	4	7	8	3	6	12	3	16	9	18	2	10
	Good	Intermediate	6	2	10	43	21	7	11	2	1	3	7	11	4	2	18	7	8	14	5	2	16
		Alternative	3	2	4	63	7	7	14	1	1	5	8	6	2	2	14	15	13	11	12	1	10
Haddock	Bad	Intermediate	15	13	19	10	8	16	18	7	7	10	5	4	8	9	7	7	8	6	6	9	7
		Alternative	15	13	19	10	8	16	18	7	7	10	5	4	8	9	7	7	8	6	6	9	7
	Good	Intermediate	14	13	13	13	14	15	17	7	7	7	7	8	8	8	7	7	7	7	7	7	7
		Alternative	14	13	13	13	14	15	17	7	7	7	7	8	8	8	7	7	7	7	7	7	7
Hake	Bad	Intermediate	4	10	4	5	8	14	55	4	9	4	3	6	12	32	2	3	4	2	2	3	12
		Alternative	8	9	8	6	11	16	44	6	7	6	4	7	11	25	4	4	6	4	4	4	9
	Good	Intermediate	9	30	2	4	23	13	19	7	26	3	6	13	14	12	1	2	4	2	4	5	2
		Alternative	8	27	3	6	20	19	17	6	22	4	7	10	14	10	2	1	7	5	7	4	2
Vermilion snapper	Bad	Intermediate	21	2	17	13	21	12	15	10	1	8	6	9	6	7	7	7	10	6	8	6	8
		Alternative	17	2	17	11	17	18	18	9	1	7	5	8	9	7	7	7	8	6	6	11	9
	Good	Intermediate	18	16	14	12	14	15	12	8	8	6	6	6	8	7	6	5	10	5	8	8	8
		Alternative	16	14	14	10	12	19	15	8	7	6	5	6	10	9	5	6	10	5	5	10	9

Note: By the end of 12 years, two 6-year blocks of posterior data are available. Across data types (7 in the first time block, 14 across time blocks 1–2), the contributions sum to 100%. Seven types of data were calculated over two 6-year time blocks, including the slope (_s), standard deviation (_v), and mean (_m) of annual catches (*C*), a fishery-independent survey index (*I*), and mean length observations (ML).

Table 6. As in Table 5, but for the delay-difference MP controlling total allowable effort (DDe).

			Data types first time block (years 1–6)							Data type first two time blocks (years 1–6, 7–12)												
			Calculated over projection years 1–6							Calculated over projection years 1–6							Calculated over projection years 7–12					
Case study	Data quality	Operating model	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m	*C*_s	*C*_v	*C*_m	*I*_s	*I*_m	ML_s	ML_m
Bluefin tuna	Bad	Intermediate	16	10	14	9	9	20	22	8	6	7	4	4	11	12	6	4	11	7	6	6	7
		Alternative	9	8	12	9	8	28	26	5	5	6	4	4	15	17	5	4	12	5	5	5	6
	Good	Intermediate	13	11	22	15	11	11	17	4	5	5	5	6	4	5	4	5	18	5	18	4	10
		Alternative	11	8	32	15	7	13	15	3	4	4	4	6	3	3	2	4	27	5	24	2	9
Cod	Bad	Intermediate	16	4	24	16	20	10	11	6	3	6	6	12	6	6	6	3	23	6	7	4	5
		Alternative	21	2	28	23	14	6	7	4	3	4	7	8	3	5	8	1	35	6	11	2	3
	Good	Intermediate	15	3	24	21	26	5	7	5	2	2	8	14	4	2	14	2	26	7	3	2	9
		Alternative	23	3	20	30	14	3	6	5	2	1	8	7	2	1	22	1	27	7	6	1	8
Haddock	Bad	Intermediate	15	13	17	10	10	16	17	8	6	9	5	5	8	10	7	7	7	7	5	8	7
		Alternative	16	11	16	11	12	17	18	8	5	8	5	6	8	9	7	7	7	7	5	9	9
	Good	Intermediate	16	13	14	12	15	14	16	7	7	7	6	8	6	6	7	7	6	6	9	7	9
		Alternative	17	9	12	13	16	18	14	8	5	5	6	8	8	7	6	5	6	6	16	6	8
Hake	Bad	Intermediate	15	14	10	8	11	19	23	15	11	3	6	4	8	8	2	8	14	4	8	4	6
		Alternative	18	12	9	9	19	18	14	14	8	3	5	7	5	4	3	10	16	4	10	4	6
	Good	Intermediate	23	17	10	10	12	17	11	14	10	2	11	4	4	2	3	17	11	7	9	4	2
		Alternative	28	14	8	10	21	12	6	14	8	3	8	7	4	2	4	14	12	6	9	5	3
Vermilion snapper	Bad	Intermediate	18	10	16	13	13	14	15	10	5	7	6	7	7	9	7	6	7	6	5	9	8
		Alternative	14	10	15	11	11	20	19	9	5	6	5	6	10	9	6	6	7	6	4	11	9
	Good	Intermediate	15	11	15	11	16	20	12	8	6	6	6	7	11	9	6	5	8	5	7	8	7
		Alternative	14	10	15	9	14	22	16	7	5	7	4	7	12	9	7	6	8	5	4	12	8

Fig. 6. As in Fig. 3, but for the indicator that (at any time the indicator was calculated) kept only those data contributing to the top 95% of the distance between the Null and Intermediate or Alternative operating models. [Colour online.]



a range of operating models that encompass plausible alternative system dynamics.

Similarly to statistical power, there were few general rules regarding which of the various data should be monitored by the indicator to detect the fishery system changes of the various case studies. A key finding, however, was that when limited to only the most contributory data, there was relatively little improvement in statistical power of the indicator. This suggests that by including a wide range of data types, there may be greater generality (can detect more types of system changes) at relatively little cost in statistical power.

In some Alternative and Intermediate operating models, even though the indicator had low statistical power, the same MP was selected for "best" performance as in the Null operating model. It is perhaps not surprising that MPs that provide sustainable fishing for both Null and Alternative operating models should lead to similar data and therefore an indicator of lower power. This further highlights the need for case-specific testing of indicators not only to identify lack of statistical power but also to establish when this is likely to be consequential. An interesting opportunity exists to establish indicators specific to operating models for which MPs are known to perform poorly. In most MSE settings, a range of operating models are proposed, and it may be the case that the best-performing MP over most operating models has a known weakness with respect to a particular operating model. Using the indicator approach developed here, we may be able to adopt the

MP but only after establishing an indicator a priori with sufficient power to detect a change towards these problematic operating model dynamics.

Each simulation of the Null and Alternative operating models are identical with the exception of the particular change in system dynamics. For example, by simulation, they include the same future recruitment strengths and observation error terms. It follows that the power analysis above could have operated on the simulation-by-simulation **difference** among posterior predicted data of the Null and Alternative models and statistically evaluated whether this differs from zero (i.e., a multivariate *t* distribution; Kotz and Nadarajah 2004). An analysis of indicators using this quantity would control perfectly for among-simulation uncertainty and have substantially higher statistical power. However, this would provide a falsely optimistic view of indicator performance, since in real applications there would be no way to know exactly which MSE simulation should be paired with the real observations. It follows that the analyses here are consistent with how the indicators would be applied in practice but understate the power to detect a real shift in dynamics if all other stochastic processes were known or controlled for.

It is typical for an MSE to be based on many operating models. It is best practice to split these into a "reference set" of operating models that include alternative hypotheses for system dynamics that address the central uncertainties and a "robustness set" that address less likely scenarios (Punt et al. 2016). The reference set is used to select a best-performing MP that can then be evaluated further by the operating models of the robustness set to determine whether performance degrades substantially over a wide range of simulated conditions. Since the MP must perform adequately across the uncertainties of the reference set, the robustness set of operating models could be used as the Alternative operating models for evaluating the power of exceptional circumstances protocols. Typically, the system dynamics of robustness operating models differ in a single property from those of a corresponding reference operating model, providing pairs of Alternative and Null operating models for statistical power analysis.

The design, data, and assumptions of the indicator could vary substantially in other applications. For example, many other types of fishery data could be used, including fishery-dependent catch rate indices, age composition data, and spatial distribution of fishing. There are many more derived quantities that characterize the magnitude and trend of such data (e.g., the ratio of small to large fish or an annual estimate of fishing rate derived by catch curve analysis of catch composition data). Providing the operating model has an appropriate observation model for the data, they can be simulated and incorporated into an indicator similar to that tested here. While we arbitrarily assumed a time-blocking of 6 years for summarizing posterior data, clearly this could be altered in other settings to better reflect the longevity of the individuals and the degree of temporal variability in stock productivity. Following previous studies, an acceptable rate of 10% false positives (alpha) was used in this indicator. In other management settings, it may be necessary to strike a trade-off between alpha and the rate of false negatives (beta), depending on management priorities and consideration of risks.

This paper did not consider operating model misspecification where both historical and future dynamics were misspecified. Clearly, the approach described here could be used to identify and test indicators in such situations. Since in these cases discrepancies among posterior data of Null and Alternative models are likely to occur over shorter projected time periods, the power of the indicator is likely to be greater than presented here.

While this paper focused on detecting temporal changes in the parameters controlling system dynamics, there are other forms of model misspecification for which similar indicators could be established. These include structural model misspecification (e.g., ignoring spatial population dynamics, use of an inappropriate stock–recruitment model), biases in observed data such (e.g., catch under-reporting), variance in observed data (e.g., unaccounted for overdispersion in catch composition data), and implementation error (e.g., violation of spatial controls or illegal catches).

While ecological indicators are commonly used to monitor trends over time and detect undesirable changes in an ecosystem (Cairns et al. 1993; Dale and Beyeler 2001), many indicators suffer from a lack of statistical power and consequently, despite expensive monitoring programs, have a low probability of detecting the negative changes (Jennings 2005; Legg and Nagy 2006). Numerous authors, in fisheries science and other disciplines, have argued that it is crucial to carry out power analysis before proposing an indicator and establishing a monitoring program (Peterman 1990; Murtaugh 1996; Hoenig and Heisey 2001; Morrison 2007). In recent years there have been an increasing number of studies that use Monte Carlo methods to evaluate the power of an indicator in detecting trends or important changes in a system (e.g., Newmark and Senzota 2003; Jennings 2005; Irvine et al. 2011). The method developed in this study differs from previous power analysis of indicators by using closed-loop simulation (MSE) to explicitly account for feedback of the management approach (the MP) on the indicator. This reveals that in some cases it may not be possible to detect significant change in the system and reject the Null operating model due to the mode of management that masks the signal of the change in the indicator. More importantly, the power analysis of indicators using MSE reveals whether the biological or ecological changes are in fact consequential to the management of the system and allows the identification of management procedures that are robust to changes in biological and ecological conditions.

Following the results of this paper, we recommend that protocols for exceptional circumstances should be tested for statistical power among the various operating models that are proposed in an MSE process. Such a power analysis could be extended to include other hypothetical scenarios for changes in system dynamics, such as those tested here. The results of this testing should focus those operating models where MP selection differs. It may be beneficial to identify those data that are particularly contributory for the indicator, but it is not advisable to limit the generality of the indicator by narrowing the types of data that it uses.

Acknowledgements

This research was supported by the David and Lucille Packard Foundation, the Natural Resources Defense Council, Fisheries and Oceans Canada, and the Marine Stewardship Council.

References

Anonymous. 2018. Report of the 2017 ICCAT bluefin stock assessment meeting. Col. Vol. Sci. Pap. ICCAT, **74**(6): 2372–2535.

Brereton, R.G. 2015. The Mahalanobis distance and its relationship to principal component scores. J. Chemom. **29**: 143–145. doi:10.1002/cem.2692.

Butterworth, D.S. 2008. Some lessons from implementing management procedures. *In* Fisheries for Global Welfare and Environment, 5th World Fisheries Congress 2008. *Edited by* K. Tsukamoto, T. Kawamura, T. Takeuchi, T.D. Beard, Jr., and M.J. Kaiser. TERRAPUB, Toyko. pp. 381–397.

Butterworth, D.S., and Punt, A.E. 1999. Experiences in the evaluation and implementation of management procedures. ICES J. Mar. Sci. **56**: 985–998. doi:10.1006/jmsc.1999.0532.

Cairns, J., McCormick, P.V., and Niederlehner, B.R. 1993. A proposed framework for developing indicators of ecosystem health. Hydrobiologia, **263**(1): 1–44. doi:10.1007/BF00006084.

Carruthers, T. 2018*a*. Operating model for East Atlantic Bluefin tuna (*Thunnus thynnus*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Bluefin_Tuna_EAtl_ICCAT/Bluefin_Tuna_EAtl_ICCAT.html [accessed April 2018].

Carruthers, T. 2018*b*. Cod 4X5Y: a DFO Case Study Operating model. Technical report for Fisheries and Oceans Canada [online]. Available from http://

www.datalimitedtoolkit.org/Case_Studies_Table/Cod_4X5Y_DFO/Cod_4X5Y_DFO.html [accessed April 2018].
Carruthers, T. 2018*c*. Operating model for Gulf of Mexico Vermillion Snapper (*Rhomboplites aurorubens*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Vermillion_Snapper_GOM_NOAA/Vermillion_Snapper_GOM_NOAA.html [accessed April 2018].
Carruthers, T.R., and Hordyk, A.H. 2018. The Data-Limited Methods toolkit (DLMtool): an R package for informing management of data-limited fisheries. Methods Ecol. Evol. **9**(12): 2388–2395. doi:10.1111/2041-210X.13081.
Carruthers, T.R., Huynh, Q., and Hordyk, A.H. 2018. Management Strategy Evaluation toolkit (MSEtool): an R package for rapid MSE testing of data-rich management procedures [online]. Available from https://github.com/tcarruth/MSEtool [accessed April 2018].
Clark, D.S., and Emberley, J. 2010. Assessment of Cod in Division 4X in 2008. DFO Can. Sci. Advis. Sec. Res. Doc. 2009/018.
Cochrane, K.L., Butterworth, D.S., De Oliveira, J.A.A., and Roel, B.A. 1998. Management procedures in a fishery based on highly variable stocks and with conflicting objectives: experiences in the South African pelagic fishery. Rev. Fish. Biol. Fish. **8**: 177–214. doi:10.1023/A:1008894011847.
Dale, V.H., and Beyeler, S.C. 2001. Challenges in the development and use of ecological indicators. Ecol. Indic. **1**: 3–10. doi:10.1016/S1470-160X(01)00003-6.
De Moor, C.L., Butterworth, D.S., and De Oliveira, J.A. 2011. Is the management procedure approach equipped to handle short-lived pelagic species with their boom and bust dynamics? The case of the South African fishery for sardine and anchovy. ICES J. Mar. Sci. **68**(10): 2075–2085. doi:10.1093/icesjms/fsr165.
DFO. 2011. Western Component (4Xopqrs5) Pollock Management Strategy Evaluation. DFO Can. Sci. Advis. Sec. Sci. Advis. Rep. 2011/054.
Dick, E.J., and MacCall, A.D. 2011. Depletion-based stock reduction analysis: a catch-based method for determining sustainable yields for data-poor fish stocks. Fish. Res. **110**: 331–341. doi:10.1016/j.fishres.2011.05.007.
DLMtool. 2018. Fishery library of documented operating models [online]. Available from http://www.datalimitedtoolkit.org/fishery_library/ [accessed April 2018].
Eigaard, O.R., Thomsen, B., Hovgaard, H., Nielsen, A., and Rijnsdorp, A.D. 2011. Fishing power increases from technological development in the Faroe Islands longline fishery. Can. J. Fish. Aquat. Sci. **68**(11): 1970–1982. doi:10.1139/f2011-103.
Garthwaite, P.H., and Koch, I. 2016. Evaluating the contributions of individual variables to a quadratic form. Austral. N.Z. J. Stat. **58**(1): 99–119. doi:10.1111/anzs.12144.
Garthwaite, P.H., and Koch, I. 2018. Partitioning of the Mahalanobis distance [online]. Available from http://users.mct.open.ac.uk/paul.garthwaite/MahalDist.html [accessed April 2018].
Gnanadesikan, R., and Kettenring, J.R. 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, **28**(1): 81–124. doi:10.2307/2528963.
Golub, G.H., and Van Loan, C.F. 1989. Matrix computations. 2nd ed. Johns Hopkins University Press, Baltimore, Md.
Harford, W.J., and Carruthers, T.R. 2017. Interim and long-term performance of static and adaptive management procedures. Fish. Res. **190**: 84–94. doi:10.1016/j.fishres.2017.02.003.
Hartigan, P.M. 1985. Computation of the dip statistic to test for unimodality. Appl. Stat. **34**: 320–325. doi:10.2307/2347485.
Hoenig, J.M., and Heisey, D.M. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am. Stat. **55**(1): 19–24. doi:10.1198/000313001300339897.
Hordyk, A. 2018*a*. Operating model for Faroe Islands Haddock (*Melanogrammus aeglefinus*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Haddock_Faroe_Islands/Haddock_Faroe_Islands.html [accessed April 2018].
Hordyk, A. 2018*b*. Operating model for Pacific Hake [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Pacific_Hake_USCan/Pacific_Hake_USCan.html [accessed April 2018].
Irvine, K.M., Dinger, E.C., and Sarr, D. 2011. A power analysis for multivariate tests of temporal trend in species composition. Ecology, **92**(10): 1879–1886. doi:10.1890/10-2138.1. PMID:22073778.
International Whaling Commission. 1999. The Revised Management Procedure (RMP) for baleen whales. J. Cetac. Res. Manage. **1**(suppl.): 251–258.
Jennings, S. 2005. Indicators to support an ecosystem approach to fisheries. Fish Fish. **6**(3): 212–232. doi:10.1111/j.1467-2979.2005.00189.x.
Kotz, S., and Nadarajah, S. 2004. Multivariate *t* distributions and their applications. Cambridge University Press.
Legg, C.J., and Nagy, L. 2006. Why most conservation monitoring is, but need not be, a waste of time. J. Environ. Manage. **78**(2): 194–199. doi:10.1016/j.jenvman.2005.04.016.
Leys, C., Klein, O., Dominicy, Y., and Ley, C. 2018. Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. J. Exp. Social Psychol. **74**: 150–156. doi:10.1016/j.jesp.2017.09.011.
Mahalanobis, P.C. 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India, **2**(1): 49–55.
Morrison, L.W. 2007. Assessing the reliability of ecological monitoring data: power analysis and alternative approaches. Nat. Areas J. **27**(1): 83–91. doi:10.3375/0885-8608(2007)27[83:ATROEM]2.0.CO;2.
Murtaugh, P.A. 1996. The statistical evaluation of ecological indicators. Ecol. Appl. **6**(1): 132–139. doi:10.2307/2269559.
NAFO. 2010. Report of the Working Group on Greenland Halibut Management Strategy Evaluation (WGMSE). NAFO/FC Doc. 11/8 [online]. Available from https://www.nafo.int/Portals/0/PDFs/fc/2011/fcdoc11-08.pdf?ver=2016-02-19-063538-810 [accessed April 2018].
NAFO. 2018. Management Strategy Evaluation — Harvest Control Rules [online]. Available from https://www.nafo.int/Science/NAFO-Frameworks/MSE.
Newmark, W.D., and Senzota, R. 2003. Power to detect trends in ecological indicators in the East Usambara Mountains, Tanzania. Afr. J. Ecol. **41**(4): 294–298. doi:10.1111/j.1365-2028.2003.00473.x.
Peterman, R.M. 1990. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. **47**(1): 2–15. doi:10.1139/f90-001.
Punt, A.E., Butterworth, D.S., De Moor, C.L., De Oliveira, J.A.A., and Haddon, M. 2016. Management strategy evaluation: best practices. Fish Fish. **17**: 303–334. doi:10.1111/faf.12104.
Rose, C., and Smith, M.D. 1996. The multivariate normal distribution. Mathematica J. **6**: 32–37.
Scott, R., Pilling, G.M., Hampton, J., Reid, C., and Davies, N. 2016. Report of the Expert Consultation Workshop on Management Strategy Evaluation. WCPFC-SC12-2016/MI-WP-05 Bali, Indonesia.
SEDAR. 2016. Stock assessment report: Gulf of Mexico vermilion snapper. Southeast Data, Assessment, and Review [online]. US National Marine Fisheries Service. SEDAR 45. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Vermillion_Snapper_GOM_NOAA/docs/S45_Final_SAR.pdf.
Stewart, I.J., Forrest, R.E., Grandin, C., Hamel. O.S., Hicks, A.C., Martel, S.J.D., and Taylor, I.G. 2011. Status of the Pacific Hake (Whiting) stock in U.S. And Canadian Waters in 2011. Prepared by the Joint Technical Committee of the U.S. and Canada Pacific Hake/Whiting Agreement, National Marine Fisheries Service and Fisheries and Oceans Canada.

Appendix A. Case studies

East Atlantic bluefin tuna

Atlantic bluefin tuna are managed by the International Commission for the Conservation of Atlantic Tunas (ICCAT) as two discrete stocks originating from natal spawning areas in the Gulf of Mexico (Western stock) and the Mediterranean (Eastern stock) (Anonymous 2018). Eastern stock assessments have estimated a large shift in the magnitude of historical recruitment after 1986, and assessment projections have included this as a principal source of uncertainty. In this case study, Null, Alternative, and Intermediate recruitment scenarios are derived from the most recent 2017 stock synthesis (Methot and Wetzel 2013) assessment (Rouyer et al. 2018). This was a conventional single-area, single stock assessment that that estimates age selectivity for 15 fleets including hook and line, bait boat, purse seine, and longline gear types (see Carruthers 2018*a* for a detailed operating model report).

Atlantic cod in NAFO area 4x5Y

Cod inhabiting the Bay of Fundy and the Scotian Shelf are managed as a discrete stock (area 4x5Y) by the North Atlantic Fishery Organization (NAFO). The stock is assessed using virtual population analysis, which estimates declines in exploitable biomass since the 1980s. These declines have occurred despite marked reductions in fishing mortality rate, thereby implying large increases in natural mortality rate (Clark and Emberley 2008; Anonymous 2016). Among other hypotheses, this increase in natural mortality rate has been attributed to rapidly increasing numbers of gray seals (*Halichoerus grypus*; see Carruthers 2018*b* for a detailed operating model report). The Null, Intermediate, and Alternative operating models for Atlantic cod mimic the VPA assessments and are standard single area, age-structured models that include the alternative scenarios for time-varying natural mortality rate.

Faroe Islands haddock

The Faroe Island longline fishery for haddock within International Council for the Exploration of the Sea (ICES) area Vb was initially managed with a set of technical measures (1986–1993) before switching to a TAC control (1994–1995) and then to an effort-regulation system (1996–2002) (Eigaard et al. 2011). Technological changes to fishing gear, including the introduction of

skewed hooks, swivel lines, and stability tanks, lead to an estimated 9% annual increase in fishing power, which undermined the effectiveness of the effort-based management system (Eigaard et al. 2011). ICES (2008) evaluated the Faroe Island haddock stock and determined that the spawning stock biomass had declined from a historical maximum of 100 000 t in 2003 to 42 000 t in 2008. We used the ICES (2008) assessment to specify an operating model for Faroe Island haddock in 1996. The Null, Alternative, and Intermediate scenarios assume stable future catchability, an average annual catchability increase of 9%, and half the observed increase (4.5%), respectively (see Hordyk 2018*a* for a detailed operating model report).

Pacific hake

The Pacific hake fishery is assessed annually by a joint technical committee of US and Canadian scientists. The 2011 assessment (Stewart et al. 2011) noted that there was a marked change in the growth pattern after 1990, with an increase in the von Bertalanffy K parameter and a corresponding decrease in the L_∞ parameter. This case study was included to investigate the influence of unobserved changes in growth on the performance of management procedures. We used the 2011 and 2017 assessments (Stewart et al. 2011; Berger et al. 2017) to condition an operating model based on knowledge of the Pacific hake stock in 1989. Three scenarios of future growth were modelled: a Null scenario assuming stable growth pattern, an Alternative scenario with the observed changes in the von Bertalanffy growth parameters, and an Intermediate scenario with half of the observed change in growth pattern (see Hordyk 2018b for a detailed operating model report).

Gulf of Mexico vermilion snapper

Alongside targeted commercial and recreational fisheries, Gulf of Mexico vermilion snapper has been subjected to heavy bycatch mortality of ages 0 and 1 fish by shrimp gear (SEDAR 2016). After the year 2000, the apical fishing mortality rate (instantaneous fishing mortality rate of the fully selected age class) of shrimp gear declined 75% from levels that were comparable to other gears combined (figure 15 in SEDAR 2016). This led to a pronounced overall shift in the age selectivity toward larger, older fish (see final figure of the detailed operating report in Carruthers 2018*c*). The Null, Alternative, and Intermediate operating model scenarios capture this shift in selectivity and were taken directly from the 2016 stock synthesis assessment, which is a conventional single area, single stock assessment model that implicitly accounts for spatial exploitation dynamics by modeling fleets by area (SEDAR 2016).

Appendix B. Management procedures

Delay-difference assessment (DD and DDe)

The delay-difference assessment methods are conditioned on fishing effort, calculated by observed annual catch divided by a fishery-independent index of abundance, and assume knife-edge selectivity at the age of maturity. Because the methods assume a fishery-independent survey, they are not sensitive to changes in catchability. The delay-difference method estimates U_{MSY}, which is used to calculate an estimate of either MSY for a TAC recommendation (DD) or E_{MSY} (effort MSY) for a total allowable effort recommendation (DDe).

Fishing selectivity matches maturity-at-length (matlenlim)

The matlenlim management procedure sets the size of retention (legal size) to be knife-edged at the estimated size of maturity. The method is sensitive to observation error in the size of maturity and, because the underlying gear selectivity pattern remains unchanged, to the degree of discard mortality on sublegal fish.

Current effort (curE)

The curE method is a simple method intended to represent a "status quo" effort-based management approach. Fishing effort for all future projection years is fixed at the level of the last historical year. The method is sensitive to variability and directional changes in fishing efficiency (i.e., catchability q), which can result in considerable changes in fishing mortality despite nominal fishing effort remaining unchanged.

Marine reserve (MRreal)

The marine reserve management procedure (MRreal) closes 10% of the unfished habitat to fishing. Effort from the closed area is reallocated to the open area so that overall effort remains unchanged for future projections (see curE above). The effectiveness of spatial closure approaches is strongly determined by the size of the closed area and the annual movement rates of the population. The operating models of these analyses are two-box models that simulate a high degree of stock mixing (nine of ten individuals move among years) among the open and closed areas. This provides spatially mixed dynamics similar to those assumed by the single area stock assessments on which the operating models are based.

Mean-catch depletion (MCD)

MCD assumes Schaefer surplus production dynamics (B_{MSY} occurs at half of carrying capacity) and that mean historical catches $\overline{C}$ are proportional to MSY. Given these assumptions, MCD aims for F_{MSY} to equal $2D\overline{C}$, where depletion D is estimated by population survey, relative abundance indices, or catch composition data. During comprehensive simulation testing, MCD generally outperformed similar "data-moderate" approaches such as depletion corrected average catch and depletion-based stock reduction analysis while requiring fewer data (Harford and Carruthers 2017). However, for average catches to be in the same order of magnitude as MSY, historical exploitation must have depleted the stock down to around B_{MSY} levels or lower.

References

Anonymous. 2016. 2016 4X5Yb Atlantic Cod (*Gadus morhua*) Stock Status Update. Canadian Science Advisory Secretariat. Science Response 2017/024.

Anonymous. 2018. Report of the 2017 ICCAT bluefin stock assessment meeting. Col. Vol. Sci. Pap. ICCAT, 74. pp. 2372–2535.

Berger, A., Grandin, C.J., Taylor, I.G., Edwards, A.M., and Cox, S. 2017. Status of the Pacific Hake (whiting) stock in U.S. and Canadian waters in 2017 [online]. Joint Technical Committee of the Pacific Hake/Whiting Agreement Between the Governments of the United States and Canada. Available from https://www.pcouncil.org/wp-content/uploads/2017/02/hake-assessment-2017-final.pdf [accessed April 2018].

Carruthers, T. 2018*a*. Operating model for East Atlantic Bluefin tuna (*Thunnus thynnus*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Bluefin_Tuna_EAtl_ICCAT/Bluefin_Tuna_EAtl_ICCAT.html [accessed April 2018].

Carruthers, T. 2018*b*. Cod 4X5Y: a DFO Case Study Operating model [online]. Technical report for Fisheries and Oceans Canada. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Cod_4X5Y_DFO/Cod_4X5Y_DFO.html [accessed April 2018].

Carruthers, T. 2018*c*. Operating model for Gulf of Mexico Vermillion Snapper (*Rhomboplites aurorubens*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Vermillion_Snapper_GOM_NOAA/Vermillion_Snapper_GOM_NOAA.html [accessed April 2018].

Clark, D.S., and Emberley, J. 2008. Assessment of Cod in Division 4X in 2008. Canadian Science Advisory Secretariat. Research Document 2009/018.

Eigaard, O.R., Thomsen, B., Hovgaard, H., Nielsen, A., and Rijnsdorp, A.D. 2011. Fishing power increases from technological development in the Faroe Islands longline fishery. Can. J. Fish. Aquat. Sci. **68**(11): 1970–1982. doi:10.1139/f2011-103.

Harford, W.J., and Carruthers, T.R. 2017. Interim and long-term performance of static and adaptive management procedures. Fish. Res. **190**: 84–94. doi:10.1016/j.fishres.2017.02.003.

Hordyk, A. 2018*a*. Operating model for Faroe Islands Haddock (*Melanogrammus aeglefinus*) [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Haddock_Faroe_Islands/Haddock_Faroe_Islands.html [accessed April 2018].

Hordyk, A. 2018*b*. Operating model for Pacific Hake [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Pacific_Hake_USCan/Pacific_Hake_USCan.html [accessed April 2018].

ICES. 2008. Report of the ICES Advisory Committee 2008. ICES Advice, Book 4. International Council for the Exploration of the Sea, Copenhagen, Denmark.

Methot, R.D., and Wetzel, C.R. 2013. Stock Synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. **142**: 86–99. doi:10.1016/j.fishres.2012.10.012.

Rouyer, T., Kimoto, A., Kell, L., Walter, J.F., Lauretta, M., Zarrad, R., Ortiz, M., Palma, C., Arrizabalaga, H., Sharma, R., Kitakado, T., and Abid, N. 2018. Preliminary 2017 stock assessment results for the Eastern and Mediterranean Atlantic bluefin tuna stock. ICCAT Col. Vol. Sci. Pap. 74. pp. 3234–3275.

SEDAR. 2016. Stock Assessment Report: Gulf of Mexico Vermilion Snapper. Southeast Data, Assessment and Review 45 [online]. Available from http://www.datalimitedtoolkit.org/Case_Studies_Table/Vermillion_Snapper_GOM_NOAA/docs/S45_Final_SAR.pdf [accessed July 2018].

Stewart, I.J., Forrest, R.E., Grandin, C., Hamel, O.S., Hicks, A.C., Martell, S.J.D., and Taylor, I.G. 2011. Status of the Pacific Hake (Whiting) stock in U.S. and Canadian Waters in 2011 [online]. Joint U.S. and Canadian Hake Technical Working Group, Final SAFE document. Available from https://www.pcouncil.org/wp-content/uploads/Pacific_Whiting_2011_Assessment.pdf [accessed April 2018].